Universitá degli Studi di Trento

Studies of the Higgs sector in $H \to ZZ \to 2\ell 2q$ and $bbH \to 4b$ semileptonic channels at CMS.

Ph.D. thesis by

Konstantin Kanishchev

Thesis advisor: Prof. Ignazio Lazzizzera

Referees:
Prof. Ugo Gasparini
Prof. Giovanni Andrea Prodi
Prof. Guido Tonelli

Facoltá di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica,

December 16, 2013

Acknowledgements

First of all, I would like to thank prof. Ignazio Lazzizzera for his supervision, resourcefulness and encouragement throughout all my endeavours during my Ph.D. school. I am also in debt to my past supervisors: Ilya Ginzburg and Maria Krawczyk, that participated in my development as physicist.

My path through the CMS collaboration was quite chaotic for the last three years and on the way I've met a lot of amazing people that were always kind and friendly towards me. I would like to thank all my colleagues from INFN CMS-Padova group. Professor Ugo Gasparini, Stefano Lacaprara, Mia Tosi, Alberto Zuchetta and Andrea Gozzelino were all enormously helpful to get me started with the basics of CMS analyses, CMS software and organization.

I am grateful HZZ2l2q group convenors Matthias Mozer and Oscar Gonzalez Lopez for taking me on board, for their guidance and expertize within the group. I'd like to give my regards to all the members of the HZZ2l2q group, especially mentioning Piergiulio Lenzi, Antonio Topiano and Valentina Gori from INFN CMS-Florence group for being determined and encouraging while we was struggling with the low-mass analysis. Also I am very grateful to Pablo Garcia, Daniel Dominguez Vazquez, Shin-Shan Yu and Yun Ju for helping me while working on various parts of the high-mass analysis.

I would also like to mention the MC expert Roberto Covarelli for guiding me through all technicalities related to centralized Monte Carlo productions. My ongoing Beyond Standard Model studies are done with Mario Pellicioni and Stefano Casasso – their feedback and constructive comments are very helpful and always welcome.

My thanks to all the members of the University of Trento physics department, to the head of Ph.D. school Professor Stefano Giorgini and the secretary Michaela Paoli for the smooth organization of the process.

Finally, I'd like to thank my family: my wife Natalia, my parents and sister for their patience and support through these years in Trento.

Contents

Chapter 1

Introduction

Particle physics is devoted to the exploration of the most fundamental constituents of matter, the way they interact and compose the world around us. As for today, it is established that all matter is made of only a dozen of fundamental particle species. These particles interact by means of four fundamental interactions: gravitational, electromagnetic, strong and weak. Three of the four fundamental interactions (with the exception of gravity) are described by the Standard Model (SM) of particle interactions.

For many years the Standard Model withstanded numerous experimental tests and provided theoretical predictions, all confirmed by the experiments at particle colliders: massive W and Z bosons were discovered at Super Proton Syncrotron, c-quark and τ-lepton were observed at Stanford Linear Accelerator and t-quark existence was confirmed by Tevatron experiments. But until 2012 one of the SM predictions stayed unconfirmed. The Standard Model predicts the existence of elementary excitation of the Higgs field – the field that is responsible for spontaneous electroweak symmetry breaking. Such excitation, called "the Higgs boson", should be a spin-0 particle that interacts with fermions and with W and Z gauge bosons.

Altough many attempts was made to observe it at various collider experiments, presence of the Higgs boson was not confirmed until 4th of July 2012. At that date, the observation of new scalar boson with mass near 125 GeV was reported by the CMS and ATLAS experiments at the Large Hadron Collider (LHC).

Despite the remarkable success of the Standard Model, there are several theoretical considerations and numerous evidence from observations in cos-

mology and astrophysics, that the SM is incomplete – some extension of the SM are needed to fully and consistently describe the Universe as we know it. There are also some hints that such extensions, called "the New Physics" or "Beyond Standard Model" theories, could be related to the electroweak sector of the Standard Model.

The thesis are devoted to my Ph.D. research activities during last three years within the CMS collaboration. My primary field of interest was the investigation of the Higgs sector of the Standard Model and its extensions.

My involvement started in the beginning of 2012 with the search of the supersymmetric Higgs, decaying in a pair of b-quarks and created in association with a $b\bar{b}$ pair (total 4 b-quarks in final state). The work was done in collaboration with CMS-Padova group. The main complication of the search was the huge multi-jet QCD background – in order to reduce it, the analysis strongly relied on the b-tagging techniques and data-driven background determination. In that work I was involved in studying and validation of various b-tagging HLT trigger paths, and evaluation of systematic errors.

After the discovery of the 125 GeV Higgs boson in the summer of 2012, another project was initiated – we attempted to detect the Higgs boson in the $H \to ZZ \to 2\ell 2b$ channel. I've participated in the feasibility studies of such analyses first working with CMS-Padova group and then with CMS-Florence group. The studied production+decay channels were:

1. $H \to ZZ \to 2\ell 2b$ with 2 leptons and 2 b-tagged jets in the final state
2. $H \to ZZ \to 2\ell 2q$ with 2leptons and 2 jets in the final state
3. $qqH \to ZZ \to 2l2q$ with Vector Boson Fusion production mode and 4 jets (2 from Higgs 2 VBF recoil jets) in the final states.

In these studies I was (to various degrees) involved in data processing on the LHC GRID, optimization of the analysis cuts, application of the machine-learning techniques to increase the analysis sensitivity. Unfortunately, all the preliminary studies had shown, that the best achievable sensitivity in all the mentioned channels were insufficient. The main reason for that was the fact that 125 GeV Higgs decay is below the ZZ decay threshold.

From the beginning of the 2013 I've continued to work on the search of the heavy Higgs ($m_H > 200$ GeV) in the $H \to ZZ \to 2\ell 2q$ channel. I was involved in estimation of systematic uncertainties, signal and background

Monte-Carlo productions. Also I was involved in the high-level statistical analysis – I've worked on interpolation of the search results and combination with searches in other channels and over 7 TeV dataset.

For several last months I was working on Beyond-Standard-Model (BSM) interpretations of the $H \to ZZ \to 2\ell 2q$ high mass search for the heavy scalar – the $m_H = 125$ GeV Higgs was discovered in a context of a whole mass-range search up to 1 TeV. Yet, after the discovery, some reinterpretations of the search results for some BSM models are possible. The two simple models for BSM are the electroweak (EWK) singlet model and the Two Higgs Doublet Model (2HDM).

Chapter 2

The Standard Model

The Standard Model (SM) of particle interactions is a product of collective work of scientists around the world throughout the last $\sim$ 50 years. It combines descriptions of electromagnetic, weak and strong interactions in a single Standard Model Lagrangian.

The theoretical foundation of the SM is the quantum field theory with $SU(3)_C \times SU(2)_T \times U(1)_Y$ gauge symmetry. The first $SU(3)_C$ multiplier is the color group associated with the strong interactions and Quantum Chromodynamics (QCD). The next two terms $SU(2)_T \times U(1)_Y$ cooresponds to the symmetry of the Glashow-Weinberg-Salam [1, 2] theory of electroweak interactions. This incorporates the spontaneous electroweak symmetry breaking (EWSB) down to electromagnetic $U(1)_{EM}$ symmetry. Just recently it was established that the mechanism behind the EWSB is the Higgs mechanism.

2.1 SM particle content

Most of the lagrangian of the SM can be deduced starting from its fermionic content. It so turns out that fermions of the Standard Model can be naturally combined in three generations $G1$, $G2$ and $G3$ as they are listed in the Table. 2.1. Subscripts ψ_L and ψ_R denote right- and left-handed chiral fields:

$$\psi_L = \frac{1-\gamma_5}{2}\psi, \quad \psi_R = \frac{1+\gamma_5}{2}\psi. \tag{2.1}$$

The Table. 2.1 presents also dimensions of representations of the color $SU(3)_C$ and isospin $SU(2)_T$ groups for each to the corresponding fermionic

$G1$	$G2$	$G3$	C	T	Y
$\begin{pmatrix}\nu_e\\ e^-\end{pmatrix}_L$	$\begin{pmatrix}\nu_\mu\\ \mu^-\end{pmatrix}_L$	$\begin{pmatrix}\nu_\tau\\ \tau^-\end{pmatrix}_L$	**1**	**2**	$-1/2$
$\begin{pmatrix}u\\ d\end{pmatrix}_L$	$\begin{pmatrix}c\\ s\end{pmatrix}_L$	$\begin{pmatrix}t\\ b\end{pmatrix}_L$	**3**	**2**	$1/6$
e_R	μ_R	τ_R	**1**	**1**	-1
u_R	c_R	t_R	**3**	**1**	$2/3$
d_R	s_R	b_R	**3**	**1**	$-1/3$
ν_{eR}	$\nu_{\mu R}$	$\nu_{\tau R}$	1	1	0

Table 2.1: Fermions of the Standard Model

field. For $U(1)_Y$ hypercharge group the value of the abelian charge Y is shown. Note that fields with different chiralities have different electroweak gauge quantum numbers which establishes the chiral nature of weak interactions.

Leptons are fermions that are singlets in $SU(3)_C$ and thus do not participate in strong interactions. *Quarks* on the other hand do have a color charge and are found in Nature as part of hadrons.

Together with all the long-approved fermionic fields of the SM Table.2.1 also shows right-handed neutrinos. Observed neutrino oscillations imply that neutrinos have non-zero masses, meaning that one can obtain right-handed neutrino from a left-handed neutrino by means of an appropriate Lorentz transformation. Still the nature of the right-handed neutrinos is not completely investigated and remains an open question in particle physics.

Denoting as ψ_f each fermionic field in Table.2.1, the complete fermionic kinetic term of the SM lagrangian simply reads:

$$\mathcal{L}_{ferm} = \bar{\psi}_f \, i\gamma^\mu D_\mu \, \psi_f, \tag{2.2}$$

where the standard convention for summation over repeated indices was adopted.

The D^μ in (2.2) is the covariant derivative that introduces vector fields compensating for the local gauge transformations of each of the symmetry groups of the SM:

$$D_\mu = \partial_\mu - ig_S G^a_\mu \mathbf{\Lambda}^a - igA^a_\mu \mathbf{T}^a - ig' B_\mu \mathbf{Y}, \tag{2.3}$$

where g_S is the strong coupling constant, g and g' are $SU(2)_T$ and $U(1)_Y$ coupling constants respectively.

The matrices $\mathbf{\Lambda}^a$, $\mathbf{T}^a$ and $\mathbf{Y}$ are operators of the adjoint representation of Lie algebra of each of the gauge groups. Gell-Mann matrices are conventionally used for $\mathbf{\Lambda}^a$ and Pauli matrices for $\mathbf{T}^a$.

The covariant derivative (2.3) introduces a number of vector gauge fields. Eight gluons G^a_μ are the strong force carriers, and four fields A^a_μ, B_μ are the would-be W,Z bosons and the photon before symmetry breaking.

Each gauge field has its associated field strength tensor which can be extracted from a commutator of covariant derivatives (2.3):

$$\left[D_\mu, D_\nu\right] = -ig_S\mathbf{\Lambda}^a G^a_{\mu\nu} - ig\mathbf{T}^a W^a_{\mu\nu} - ig'\mathbf{Y}B_{\mu\nu}. \tag{2.4}$$

Extracting the field strength tensors, one gets:

$$\begin{aligned} G^a_{\mu\nu} &= \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_S f^{abc} G^b_\mu G^c_\nu \\ W^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g\varepsilon^{abc} W^b_\mu W^c_\nu, \\ B_{\mu\nu} &= \partial_\mu B^a_\nu - \partial_\nu B^a_\mu \end{aligned} \tag{2.5}$$

where ε^{abc} and f^{abc} are the structure constants of $SU(2)$ and $SU(3)$ groups.

The kinetic terms of the gauge fields is then written in the following form:

$$\mathcal{L}_{gaug} = -\frac{1}{4}G^a_{\mu\nu}G^{a,\mu\nu} - \frac{1}{4}W^a_{\mu\nu}W^{a,\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{2.6}$$

2.2 The Higgs mechanism

The lagrangian terms (2.2) and (2.6) still fail to adequately describe the Nature. While we observe nonzero masses of leptons and quarks and, also, the existence of three massive vector bosons was established, the lagrangian however lacks any mass terms in the description.

It furthermore turns out that introduction of such terms both for fermions and for gauge boson faces certain difficulties:

- **For vector bosons** mass terms $m^2 A_\mu A^\mu$ are of the degree $D = -\frac{3}{2}$ and, therefore, non-renormalizable. Introduction of such terms will challenge the consistency of the whole gauge theory basis of the model.

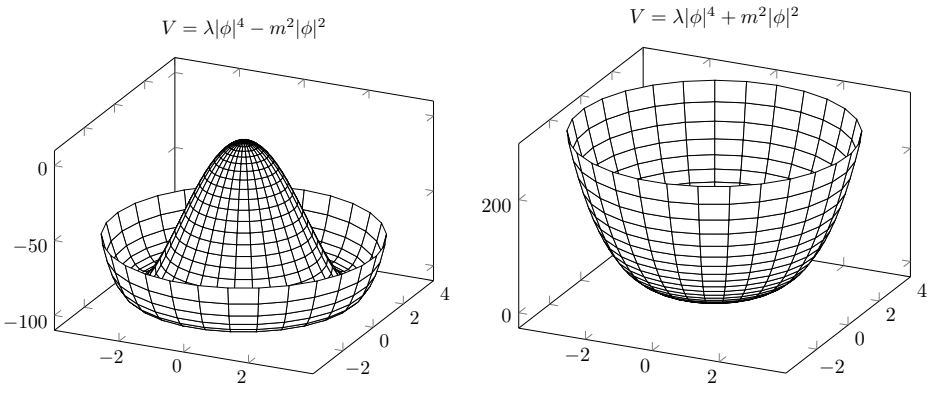


Figure 2.1: Higgs mechanism

- **Dirac fermion** mass term reads $m\bar{\psi}\psi = m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$. One can check that he cannot construct such gauge-invariant term using fields from Table. 2.1.

These obstacles can be circumvented [3–5] by introduction of an extra scalar $SU(2)_T$ doublet with hypercharge $Y = 1/2$, called the Higgs field:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{2.7}$$

The general from for the gauge-invariant renormalizable lagrangian for such a field would be:

$$\mathcal{L}_{higgs} = (D_\mu\phi)^\dagger(D_\mu\phi) - \left(\lambda(\phi^\dagger\phi)^2 - m^2(\phi^\dagger\phi)\right). \tag{2.8}$$

The negative sign of the quadratic term in (2.8) is essential for the phenomenon of the spontaneous symmetry breaking (see. Fig2.1). The field ϕ obtains nonzero vacuum expectation value (v.e.v.), which can be transformed into the following form by a global $SU(2)_T$ rotation:

$$\langle\phi\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.9}$$

The value $v = m/\sqrt{2}$ is the minimum of the Higgs potential. Considering small fluctuations around that vacuum, we get:

$$\phi(x) = \begin{pmatrix} G^+(x) \\ \dfrac{v + h(x) + iG^0(x)}{\sqrt{2}} \end{pmatrix}. \tag{2.10}$$

The fields G^0, G^+ (and its conjugate G^-) are the so-called Goldstone modes, appearing in models with the spontaneous symmetry breaking. Goldstone modes are massless in absence of gauge symmetries. The essence of the Higgs mechanism lays in the observation that when one combines massless Goldstone modes with massless gauge fields, the Goldstone modes become the longitudinal degrees of freedom of the gauge fields giving rise to the mass terms of the combined vector bosons.

The simplest way to see this is to work in the so-called unitarity gauge [6]. The gauge is chosen in such a way that all the Goldstone modes in (2.10) are cancelled by an appropriate local gauge transformation:

$$\phi(x) = e^{-i\rho(x)}e^{-i\vec{\tau}\vec{\theta}(x)} \times \begin{pmatrix} G^+(x) \\ \dfrac{v + h(x) + iG^0(x)}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{2.11}$$

The kinetic term of the Higgs lagrangian (2.8) is then can be written as:

$$(D_\mu\phi)^\dagger(D_\mu\phi) = \frac{1}{2}\left|(\partial_\mu - igA^a_\mu\mathbf{T}^a - ig'B_\mu\mathbf{Y}) \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}\right|^2. \tag{2.12}$$

Substituting the hypercharge value for the Higgs and expanding isospin matrix we obtain:

$$\begin{aligned} &= \frac{1}{8}\left|\begin{pmatrix} 2\partial_\mu - i(gA^3_\mu + g'B_\mu) & -i\sqrt{2}gA^+_\mu \\ i\sqrt{2}gA^-_\mu & 2\partial_\mu + i(gA^3_\mu - g'B_\mu) \end{pmatrix} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}\right|^2 \\ &= \frac{1}{8}\left|\begin{pmatrix} -i\sqrt{2}gA^+_\mu(v + h) \\ 2\partial_\mu h + i(gA^3_\mu - g'B_\mu)(v + h) \end{pmatrix}\right|^2 \end{aligned} \tag{2.13}$$

Expanding the square and leaving only kinetic and mass terms :

$$= \frac{1}{2}\partial_\mu h\,\partial^\mu h + \frac{g^2}{4}v^2W^+_\mu W^{-\mu} + \frac{g^2 + g'^2}{8}v^2Z_\mu Z^\mu + \cdots \tag{2.14}$$

Where the massive $W_\mu^\pm$ and Z_μ vector boson states and the massless photon state A_μ are introduced:

$$\begin{aligned} W_\mu^\pm &= \frac{1}{\sqrt{2}}(A_\mu^1 \pm A_\mu^2) \qquad & m_W &= \frac{v}{2}g \\ Z_\mu &= \frac{1}{\sqrt{g^2+g'^2}}(gA_\mu^3 - g'B_\mu) \qquad & m_Z &= \frac{v}{2}\sqrt{g^2+g'^2} \\ A_\mu &= \frac{1}{\sqrt{g^2+g'^2}}(g'A_\mu^3 + gB_\mu) \qquad & m_A &= 0 \end{aligned} \tag{2.15}$$

The non-renormalizability of the massive vector bosons is solved because the original lagrangian (before SSB) doesn't contain non-renormalizable terms. The appearance of a non-vanishing v.e.v. of the Higgs field does not affect the asymptotic behavior of the QFT correlation functions, thus preserving the renormalizability of the model.

2.3 Yukawa sector of the SM

We are left with the task of assigning masses to fermions from the Table. 2.1 – the fields alone cannot be combined in gauge-invariant Dirac mass terms. The issue can also be resolved using the same $Y = 1/2$ Higgs field (2.7) with non-vanishing vacuum expectation value.

2.3.1 Quark mass terms and CKM matrix

Let us denote, for brevity, all the quark fields as $u_{L,R}^i$ and $d_{L,R}^i$, with index i indicating the generation to which the fermion belongs. The most general gauge-invariant renormalizable lagrangian for Higgs-quark interaction can then be written in the following compact form:

$$-\mathcal{L}_Y = \Gamma'_{ij}\begin{pmatrix}\bar{u}_L^i \\ \bar{d}_L^i\end{pmatrix}\phi\, d_R^j + \Delta'_{ij}\begin{pmatrix}\bar{u}_L^i \\ \bar{d}_L^i\end{pmatrix}\tilde{\phi}\, u_R^j + h.c. \tag{2.16}$$

Here $\tilde{\phi} = \varepsilon^{ab}\phi_b^\dagger$ is the Higgs field in conjugate T and Y representations. After we perform the untarity gauge transformation (2.11), only one component of ϕ remain nonzero:

$$-\mathcal{L}_Y = \Gamma'_{ij}\begin{pmatrix}\bar{u}_L^i \\ \bar{d}_L^i\end{pmatrix}\cdot\begin{pmatrix}0 \\ \frac{v+h}{\sqrt{2}}\end{pmatrix} d_R^j + \Delta'_{ij}\begin{pmatrix}\bar{u}_L^i \\ \bar{d}_L^i\end{pmatrix}\cdot\begin{pmatrix}\frac{v+h}{\sqrt{2}} \\ 0\end{pmatrix} u_R^j + h.c. \tag{2.17}$$

The matrices Γ' and Δ' are not necessary real or even hermitian. Each of these matrices can be diagonalized using two unitary matrices:

$$\Gamma' \to U_\gamma \, \Gamma \, W_\gamma^\dagger \qquad \Delta' \to U_\delta \, \Delta \, W_\delta^\dagger$$

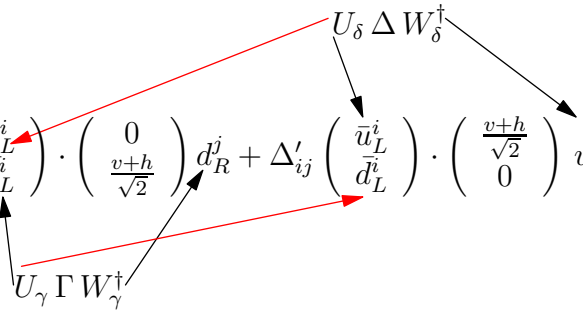
The transformed matrices Γ and Δ are diagonal. So what one should do to obtain mass-eigenstate genration is do choose an appropriate basis in space of families by performing appropriate rotations for individual fermionic fields:

$$\begin{aligned} d_R &\to W_\gamma \, d_R \qquad & u_R &\to W_\delta \, u_R \\ d_L &\to U_\gamma \, d_L \qquad & u_L &\to U_\delta \, u_L \end{aligned} \tag{2.18}$$

The diagonal elements of the Yukawa matrices will result in flavor-diagonal mass terms of the form $v_i \Gamma_i (\bar{u}_L^i u_R^i + h.c.)$ with masses of individual fermions naturally appearing as $m_u^i = v\Gamma^i$.

Consider now how the rotation in generation-space affect the fermionic kinetic term (2.2). The derivative $\bar{\psi} \partial_\mu \gamma^\mu \psi$ stays diagonal after the unitary transformations (2.18). But notice, that upper and lower components of a fermionic isospinor are changed *separately* by U_δ and U_γ:

$$-\mathcal{L}_Y = \Gamma'_{ij} \begin{pmatrix} \bar{u}_L^i \\ \bar{d}_L^i \end{pmatrix} \cdot \begin{pmatrix} 0 \\ \frac{v+h}{\sqrt{2}} \end{pmatrix} d_R^j + \Delta'_{ij} \begin{pmatrix} \bar{u}_L^i \\ \bar{d}_L^i \end{pmatrix} \cdot \begin{pmatrix} \frac{v+h}{\sqrt{2}} \\ 0 \end{pmatrix} u_R^j$$



The transformation won't produce generation-diagonal terms for interactions that change isospin. The corresponding term in the expansion of the covariant derivative reads:

$$g W_\mu^+ \cdot \bar{u}_L^i \gamma^\mu (U_\delta^\dagger U_\gamma)_{ij} d_L + h.c. \tag{2.19}$$

The matrix $V_{ij} = (U_\delta^\dagger U_\gamma)_{ij}$ is called "the Cabibbo Kobayashi Maskawa" (CKM) matrix [7, 8] for quark mixing. It is responsible for generation-mixing charged currents and CP-violation in the Standard Model.

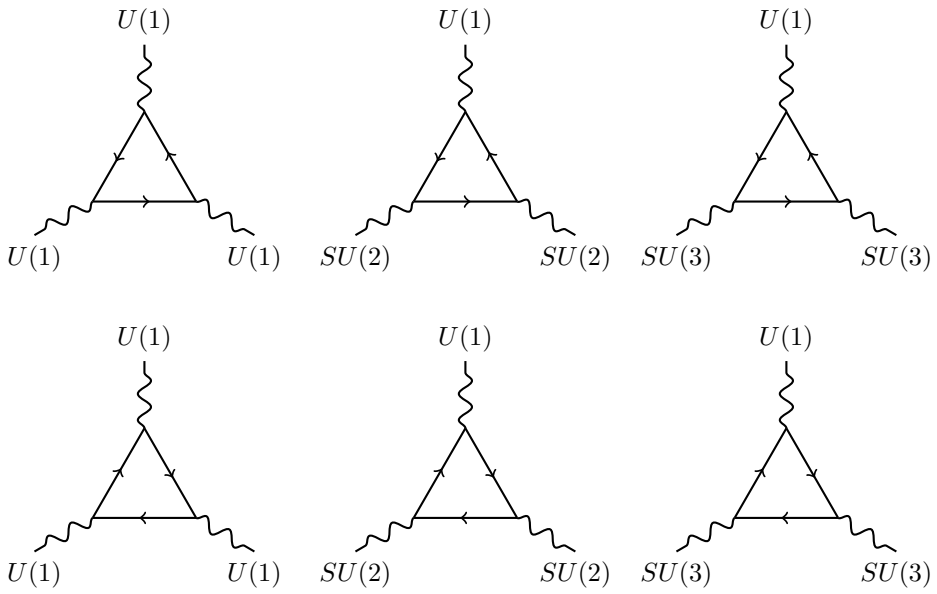


Figure 2.2: Triangle diagrams that contribute to gauge anomalies of GWS theory.

2.3.2 Leptons and anomaly cancellation

Leptonic mass terms are introduced in a similar fashion. Discarding, for now, the right-handed neutrinos and following the outlined above procedure we get the leptonic Yukawa terms:

$$-\mathcal{L}_Y = \Theta'_{ij} \begin{pmatrix} \bar{\nu}^i_L \\ \bar{\ell}^i_L \end{pmatrix} \cdot \begin{pmatrix} 0 \\ \frac{v+h}{\sqrt{2}} \end{pmatrix} \ell^j_R + h.c. \tag{2.20}$$

The Yukawa matrix Θ'_{ij} is diagonalized in the same way, producing masses for the three leptons e, μ and τ, leaving neutrinos massless. Since the neutrino mass terms are absent (2.20), such construction does not produce any lepton-flavor charged currents.

For quite a long time, the neutrinos were believed to be massless, however, now after the observation of neutrino oscillations it became clear that neutrinos do have masses and that neutrinos' weak interaction eigenstates are quite different from their mass eigenstates – just as it happens for quarks.

The corresponding mixing matrix, analogous to CKM matrix, is called the Pontecorvo–Maki–Nakagawa–Sakata (PMNS) matrix.

Curious observation, named "Quark-Lepton complementarity" [9, 10], appears from studying both PNMS and CKM matirces: the mixing angles, corresponding to non-diagonal terms in both matrices, seem to add roughly to $\pi/4$. Such a "diagonal" relation might be viewed as a hint for some deeper symmetry between quarks and leptons realized by means of some GUT model.

Another important feature also related to mutual participation of quarks and leptons is the phenomenon of *anomaly cancellation*. Anomaly is a breakdown of some symmetry of classical lagrangian for a Quantum Field Theory after accounting for radiative corrections. Soon after the Glashow-Weinberg-Salam theory was introduced, it was realized [11] that in general non-abelian chiral models suffer from *chiral anomalies*.

To see this, consider the triangle diagrams on Fig. 2.2. The diagrams might potentially generate an axial-current term, which would break gauge symmetry and render the theory non-renormalizable. The right- and left-handed fermionic loops would have canceled each other if the bosons won't distinguish between them. Since this is not the case in chiral models, number of non-trivial cancellations must occur for the theory to be consistent.

For example the contribution from two loops, involving three $U(1)$ gauge bosons (first column on Fig. 2.2) will be proportional to the following sum over fermionic fields:

$$\begin{aligned} &3\sum_q \gamma^5 \mathbf{Y}^3 + \sum_\ell \gamma^5 \mathbf{Y}^3 = \\ &= 3\left(2\left(\frac{1}{6}\right)^3 - \left(\frac{2}{3}\right)^3 - \left(-\frac{1}{3}\right)^3\right) + 2\left(-\frac{1}{2}\right)^3 - (-1)^3 = 0, \end{aligned} \tag{2.21a}$$

where factor 3 represents the color degree of freedom of the quarks and γ^5 sets the opposite signs of left and right-handed contributions

The second and third pair of diagrams on Figure. 2.2 depend on gauge indexes a and b. The anomaly must vanish for any values of the isospin and color indexes. Miraculously, it does also happen:

$$\sum_L \left[\mathbf{T}^a, \mathbf{T}^b\right] \mathbf{Y} = \frac{1}{2}\delta^{ab} \sum_L \mathbf{Y} = \delta^{ab}\left(-\frac{1}{2} + 3\,\frac{1}{6}\right) = 0, \tag{2.21b}$$

$$\sum_q \left[\mathbf{\Lambda}^a, \mathbf{\Lambda}^b\right] \mathbf{Y} = \frac{1}{2}\delta^{ab} \sum_q \mathbf{Y} = \frac{1}{2}\delta^{ab} \cdot \left(2\,\frac{1}{6} - \frac{1}{3}\right) = 0, \tag{2.21c}$$

respectively for $SU(2)$ left-handed fermions and $SU(3)$ quarks.

The Glashow-Weinberg-Salam theory is therefore a chiral non-abelian gauge theory, which is safe from axial anomalies because of the cancellation between quark and lepton contributions. For such cancellation to occur, the fermions must come in a "generation" multiplets, as it is indicated in the Table 2.1.

Bibliography

[1] Sheldon L. Glashow. "Partial-symmetries of weak interactions". *Nuclear Physics* 22.4 (1961), pp. 579 –588. ISSN: 0029-5582.

[2] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. "Broken Symmetries". *Phys. Rev.* 127 (3 1962), pp. 965–970.

[3] P.W. Higgs. "Broken symmetries, massless particles and gauge fields". *Physics Letters* 12.2 (1964), pp. 132 –133. ISSN: 0031-9163.

[4] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". *Phys. Rev. Lett.* 13 (9 1964), pp. 321–323.

[5] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". *Phys. Rev. Lett.* 13 (20 1964), pp. 585–587.

[6] Steven Weinberg. "General Theory of Broken Local Symmetries". *Phys. Rev. D* 7 (4 1973), pp. 1068–1082.

[7] Nicola Cabibbo. "Unitary Symmetry and Leptonic Decays". *Phys. Rev. Lett.* 10 (12 1963), pp. 531–533.

[8] Makoto Kobayashi and Toshihide Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". *Progress of Theoretical Physics* 49.2 (1973), pp. 652–657.

[9] Robert Foot and Henry Lew. "Quark-lepton-symmetric model". *Phys. Rev. D* 41 (11 1990), pp. 3502–3505.

[10] Bhag C. Chauhan et al. "Quark-lepton complementarity, neutrino and standard model data predict theta ($\theta_{13}^{PMNS} = 9^{+1}_{-2})^{\circ}$". *Eur.Phys.J.* C50 (2007), pp. 573–578. arXiv:`hep-ph/0605032 [hep-ph]`.

[11] David J. Gross and R. Jackiw. "Effect of Anomalies on Quasi-Renormalizable Theories". *Phys. Rev. D* 6 (2 1972), pp. 477–493.

Chapter 3

Beyond Standard Model

Despite its enormous success, the Standard Model leaves several unresolved questions in our understanding of the Nature. Such questions motivate theorists to construct models (the so-called Beyond-Standard Model theories) that extend the SM in various ways.

In this chapter we first briefly review some of the motivations for BSM, emphasizing all the hints for the New Physics to be linked to electroweak sector of the SM. Then we will review three BSM models, relevant for the present work: scalar electroweak (EWK) singlet model, the Two Higgs Doublet Model (2HDM) and the minimal supersymmetric extension of the SM (MSSM).

3.1 Unresolved issues of the Standard Model

3.1.1 Dark Matter

Nowadays, an overwhelming evidence from astrophyisics observations suggests that apart from all known types of matter (and energy in form of electromagnetic radiation) there are extra ingredients, building up the Universe. In fact, all the existing evidence suggests that the "ordinary" baryonic matter amounts to only about 5% of the entire Universe's content (which is, many physicists admit, is simultaneously embarrassing and exciting).

According to the prevailing cosmological ΛCDM model, the rest 95% of the Universe is made of two types of "Dark" constituents:

- **Dark Matter** (accounting for 27% of the Universe) is seen to affect galaxy rotation curves and dynamics of galaxy clusters. Evidence for existence of the Dark Matter (DM) are also provided by gravitational lensing observations, by simulations of large-scale structure formation of the Universe and from studies of anisotropy of the Cosmic Microwave Background radiation.

- **Dark Energy** (accounting for the rest 68%) is a hypothetical form of energy, contributing to the acceleration of the Universe's expansion. The accelerated rate of expansion of the Universe was discovered in studies of Type Ia supernovae distribution. Another confirmation for the existence of the of Dark Energy comes from studies of its "footprints" in CMB anisotropies.

The dominant theoretical explanation for Dark Matter is the Weakly Interacting Massive Particle or "WIMP" hypothesis – the DM is conjectured to be made of quite massive and, therefore, cold (following the classical Boltzman thermal distribution) particles, that do not participate in electromagnetic or strong interactions.

Obviously, WIMPs do participate in gravitational interaction, but more importantly they are also assumed to be weakly interacting. Primary motivation for that assumption is the so-called "WIMP miracle": During cooling of the Early Universe the DM particles and antiparticles were first in thermal equilibrium. At temperatures lower than the particle mass they started to co-annihilate into SM particles. The co-annihilation stopped at the "freeze out" point when the annihilation rate was overpowered by the Hubble expansion rate. The present Dark Matter abundance from this process can be estimated [1] by equating the Hubble expansion and annihilation rate:

$$\Omega_{DM}h^2 \sim x^{3/2}\left(\frac{30\,\text{fb}}{\sigma}\right), \quad x = \frac{m}{T} \sim 16.3 + \ln\left[\left(\frac{\sigma}{10\,\text{fb}}\right)\left(\frac{m}{\text{GeV}}\right)\right] \tag{3.1}$$

The alleged "miracle" occurs when one uses the weak-scale masses GeV and cross-sections: $m \sim 100$ GeV and $\sigma \sim 10$ fb in the (3.1) expression. The obtained density value is around the observed value of Dark Matter density. It seems natural to assume, therefore, that this is not just a coincidence, but the Dark Matter is somehow related to Electroweak physics, extending our models beyond SM.

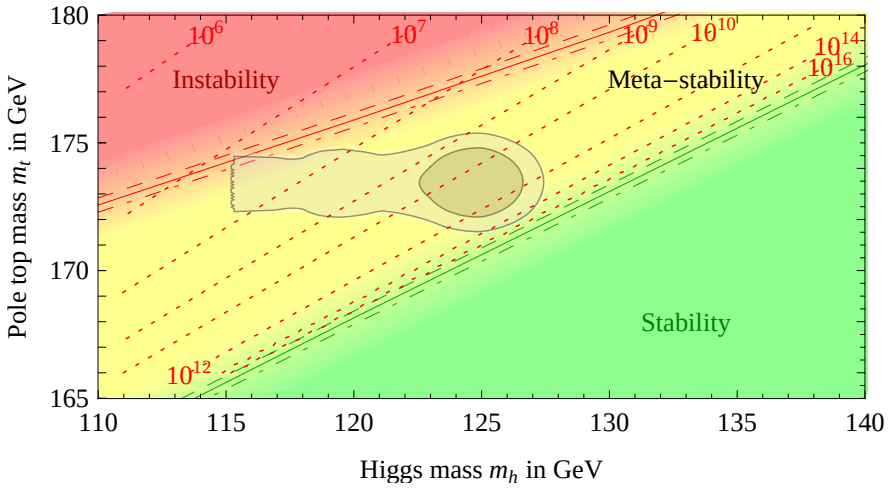


Figure 3.1: Contour lines denote the instability scale in GeV.

3.1.2 Vacuum energy and stability

To account for the Dark Energy contribution, uniform (or very slowly varying) energy density is introduced in cosmological models by means of the famous Cosmological Constant Λ-term in the in the Einstein's equations:

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}. \tag{3.2}$$

It seems tempting to associate this vacuum energy with energy of the QFT vacuum of particle physics. It so happens, however, that the measured value of Λ obtained from the CMB anisotropy measurement [2] is many orders of magnitude smaller than the one obtained from QFT [3] – the discrepancy factor of 10^{120} makes this problem "the worst fine-tuning problem in physics".

Another cosmological problem, associated with ground state of the Higgs potential is the fact that the $m_H = 126$ GeV Higgs boson violates [4] the so-called stability bound.

It turns out after calculations of the renormaliztion group running of the Higgs self-coupling λ in (2.8), that it becomes negative at some at some higher scales. This means that the Higgs potential becomes unstable at that scale, makin the theory inconsistent. Figure. 3.1 shows the stability, instability and meta-stability regions in $m_h \times m_t$ plane. As one can see, the measured values favor strongly the meta-stability region.

Whether the metastable region is acceptable is somewhat debatable. As a matter of fact, the estimated lifetime of such vacuum could be many orders of magnitude larger than the age of the Universe; still the explanation is needed for the Universe to be started in a configuration whose energy is very far from the minimum, in a metastable valley.

3.1.3 Naturalness scale of the SM

Figure 3.2 shows the 1-loop Standard Model contributions to the Higgs boson self-energy. One can see that every contribution is quadratically divergent. Assuming that SM is correct up to a scale Λ (and assuming that all the diagrams on Fig. 3.2 break roughly at the same scale), the combined 1-loop radiative correction reads:

$$\delta m_H^2 = \frac{3\Lambda^2}{8\pi^2 v^2}\left[2m_W^2 + m_Z^2 + m_H^2 - 4m_t^2\right] \sim -\frac{\Lambda^2}{25} \tag{3.3}$$

It follows from (3.3) that if SM is correct up to energies of, say $\Lambda = 5$ TeV, then the observed physical Higgs boson mass $m_H^2 = 126^2\,\mathrm{GeV}^2$ will require a cancellation between bare Higgs mass $m_{H,bare}^2$ and the correction δm_H^2. Both of which should be 100 times larger than m_H^2.

Whether the two order of magnitude cancellation between $m_{H,bare}^2$ and δm_H^2 is abnormal at $\Lambda = 5$ TeV could be still a matter of taste. But the quadratic dependence of the divergence on Λ quickly makes apparent the fact that such cancellation requires some really unnatural fine-tuning of the SM parameters at even higher scales.

To solve that problem (usually called "*the hierarchy problem*") we would require some extra BSM New Physics at scales $\Lambda \simeq 1 \div 5$ TeV, linked to the electroweak sector.

3.1.4 Origin of neutrino masses

The question of origin of neutrino masses was briefly mentioned in the previous chapter. The SM procedure of introduction of mass terms via gauge-invariant Yukawa interactions with Higgs field should in principle work also for neutrinos. The very small neutrino masses as compared to masses of quarks and leptons then require and explanation.

The Yukawa terms require us to have both left-handed and right-handed fermions. The right-handed counterpart of the observed left-handed neutri-

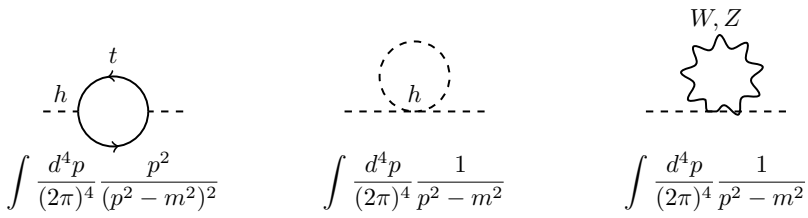


Figure 3.2: 1-loop diagrams, contributing to the Higgs self-energy.

nos is shown as the last "grayed" line on the Table. 2.1. As one can see ν_R is a singlet in all gauge groups (the "sterile" neutrino). This means that there is one more term that is permitted by renormalizability and gauge-invariance:

$$M\,\bar{\nu}_R \nu_R^\dagger \tag{3.4}$$

The mass M would then be the only dimensional parameter that enter the SM lagrangian apart from the Higgs mass. Therefore it in principle could be related to some high New Physics scale: $M \sim M_{GUT}$. Expanding the Dirac bispinors we get:

$$\begin{pmatrix} \bar{\nu}_L & \bar{\nu}_R^\dagger \end{pmatrix} \begin{pmatrix} 0 & m_{\nu EW} \\ m_{\nu EW}^T & M \end{pmatrix} \begin{pmatrix} \nu_L \\ \bar{\nu}_R^\dagger \end{pmatrix} \tag{3.5}$$

Diagonalization of this matrix naturally generates very small mass egienstates $m_\nu \sim m_{EW\nu}^2/M$. Setting $M \sim M_{GUT} \sim 10^{15}$ GeV and $m_{\nu EW} \sim 100$ GeV we obtain $m \simeq 0.01$ eV, which roughly corresponds to neutrino mass scales.

The outlined construction is the simplest implementation of the so-called "seesaw" mechanism used for explanation of relative sizes of neutrino masses. Different types of seesaw model produce various extensions of the SM, linked to other BSM models and usualy related to Electroweak sector.

3.2 Electroweak singlet

Electroweak singlet models [5] are simple models that introduce a hidden sector to the SM – some number of (usually unspecified) fields that almost completely decoupled from the usual SM sector. For such a "shadow" sector

to have some experimentally observable effects there must exist at least some couplings to the SM fields.

Very simple and yet natural way of realizing that program is to to introduce an extra complex scalar singlet ϕ_H, which is invariant under $U(1)_{hid}$ gauge symmetry. There are two types of renormalizable interactions that could be exploited to connect the hidden sector to the SM.

First possibility is connecting the hidden gauge $U(1)_{hid}$ and the SM hypercharge $U(1)_Y$ symmetry through the $\sim B_{\mu\nu}C^{\mu\nu}$ kinetic term, with $B_{\mu\nu}$ and $C_{\mu\nu}$ being the field strengh tensors as in (2.5) and (2.6). That approach leads to the so-called Z' phenomenology [6] and will not be considered in the present work.

Another way to connect "shadow" and SM sectors with the renormalizable interactions [5] is to introduce a gauge-invariant coupling to the Standard Model Higgs field ϕ_{SM} through $\sim (\phi_{SM}^\dagger \phi_{SM})|\phi_H|^2$ term. The general form of the scalar sector lagrangian is then:

$$\begin{aligned}\mathcal{L} = &|D_\mu \phi_{SM}|^2 - \lambda(\phi_{SM}^\dagger \phi_{SM})^2 + m_{SM}^2(\phi_{SM}^\dagger \phi_{SM}) \\ &+ |D_\mu \phi_H|^2 - \lambda'|(\phi_H|^4 + \mu_H^2|\phi_H|^2 + \lambda_{12}(\phi_{SM}^\dagger \phi_{SM})|\phi_H|^2.\end{aligned} \tag{3.6}$$

One can see from (3.6) that it is possible that the extra scalar obtains some vacuum expectation value $\langle \phi_H \rangle = \xi$ if $m_H^2 > 0$. When $m_H^2 < 0$ the vacuum expectation is $\xi = 0$ and the Z_2 ($\phi_H \to -\phi_H$) symmetry of the lagrangian isn't broken. This means that the corresponding massive scalar particles can only be pair-produced at colliders, leading to a parity-conserving dark matter phenomenology.

Still, the primary motivation for such models is to have a spontaneously broken gauge symmetry in the hidden sector, which, as it will be described further, leads to a simple "SM-like" phenomenology at LHC. Decomposing around the vevs of both scalars $\langle \phi_{SM} \rangle = v$, $\langle \phi_H \rangle = \xi$ we get:

$$\phi_{SM} = \frac{1}{\sqrt{2}} \begin{pmatrix} G^\pm \\ v + \eta_{SM} + iG^0 \end{pmatrix}, \quad \phi_H = \frac{1}{\sqrt{2}} \left(\xi + \eta_H + iG'\right), \tag{3.7}$$

where $G^\pm$ and G^0 are the SM Goldstone modes, G' is the Goldstone mode for the hidden sector, which provides mass for Z' boson. The massive degrees of freedom η_{SM} and η_H are mixed. The corresponding mass matrix is:

$$M = \begin{pmatrix} 2\lambda v^2 & \lambda_{12} v\xi \\ \lambda_{12} v\xi & 2\lambda' \xi^2 \end{pmatrix}. \tag{3.8}$$

The physical states and their masses are then obtained as the eigenvectors and eigenvalues of the mass matrix. The eigenvectors can be represented as a 2D rotation in the $\{\eta_{SM}, \eta_H\}$ space with a certain angle ω:

$$\begin{pmatrix} h \\ H \end{pmatrix} = \begin{pmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{pmatrix} \begin{pmatrix} \eta_{SM} \\ \eta_H \end{pmatrix}, \tag{3.9}$$

$$m^2_{h,H} = \lambda v^2 + \lambda' \xi^2 \pm \sqrt{(\lambda v^2 - \lambda' \xi^2)^2 + \lambda_{12} v^2 \xi^2}. \tag{3.10}$$

Further we will assume that h is the Standard Model Higgs with $m_h = 126$ GeV and that $m_H > m_h$.

Due to the mixing (3.9) both h and H are coupled to the SM sector and to the hidden sector. Strengths of the corresponding couplings is determined by the values of $c_\omega = \cos\omega$ and $s_\omega = \sin\omega$ correspondingly. Consequently the SM production cross-sections and SM width of these scalars are simply scaled as squares of these multipliers:

$$\sigma(SM \to h) = c^2_\omega \sigma^{SM}, \quad \Gamma(h \to SM) = c^2_\omega \Gamma^{SM}, \tag{3.11a}$$

$$\sigma(SM \to H) = s^2_\omega \sigma^{SM}, \quad \Gamma(H \to SM) = s^2_\omega \Gamma^{SM}, \tag{3.11b}$$

where σ^{SM} and Γ^{SM} are cross-sections and partial widthes of the Standard Model Higgs boson. E.g. in case when there is only the SM lagrangian without the extra sector (or when $c_\omega = 1$ and the hidden sector is completely decoupled from SM).

The relations (3.11) imply that the EWK singlet model phenomenology is very similar to the heavy-Higgs SM phenomenology. That makes the EWK singlet model very attractive from phenomenological point of view in the "post-Higgs discovery times". Reinterpretation of the results of the heavy Higgs searches is very straightforward in this model – one just needs to rescale cross-sections and couplings to get the exclusions on the model.

3.3 2HDM

The 2HDM presents the simplest extension of the standard Higgs mechanism with two scalar doublets with $Y = 1$. This model offer a natural mechanism for spontaneous CP violation [7] and has a number of interesting cosmological consequences [8].

As in the SM, the lagrangian of the model contains a Higgs and Yukawa sectors, plus fermionic and gauge kinetic terms:

$$\mathcal{L}_{2HDM} = \mathcal{L}_h + \mathcal{L}_Y + \mathcal{L}_f + \mathcal{L}_g. \tag{3.12}$$

The fermionic content and the gauge group of the model stay unchanged. So $\mathcal{L}_f$ and $\mathcal{L}_g$ are exactly the same as in the SM. The differences between 2HDM and SM arise from the second Higgs doublet, which have the same quantum number as the one in minimal SM:

$$\phi_1 = \begin{pmatrix} \phi_{11} \\ \phi_{12} \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} \phi_{21} \\ \phi_{22} \end{pmatrix}. \tag{3.13}$$

The Higgs sector of the 2HDM lagrangian then reads:

$$\mathcal{L}_h = D_\mu \phi_1^\dagger D^\mu \phi_1 + D_\mu \phi_2^\dagger D^\mu \phi_2 - V. \tag{3.14}$$

With the most general form of the 2HDM, the potential is

$$\begin{aligned} V = {} & \frac{\lambda_1}{2}(\phi_1^\dagger \phi_1)^2 + \frac{\lambda_2}{2}(\phi_2^\dagger \phi_2)^2 + \lambda_3(\phi_1^\dagger \phi_1)(\phi_2^\dagger \phi_2) + \lambda_4(\phi_1^\dagger \phi_2)(\phi_2^\dagger \phi_1) \\ & + \left[\frac{\lambda_5}{2}(\phi_1^\dagger \phi_2)^2 + \lambda_6(\phi_1^\dagger \phi_1)(\phi_1^\dagger \phi_2) + \lambda_7(\phi_2^\dagger \phi_2)(\phi_1^\dagger \phi_2) + \text{h.c.}\right] \\ & - \frac{1}{2}\left[m_{11}^2(\phi_1^\dagger \phi_1) + m_{22}^2(\phi_2^\dagger \phi_2) + m_{12}^2(\phi_1^\dagger \phi_2) + m_{12}^{2*}(\phi_2^\dagger \phi_1)\right]. \end{aligned} \tag{3.15}$$

Here λ_{1-4}, m_{11}^2 and m_{22}^2 are real and λ_{5-7}, m_{12}^2 are generally complex parameters. So the 2HDM in the most general formulation requires 14 real parameters to fully establish the model.

3.3.1 2HDM vacuum structure

After the symmetry breaking, fields ϕ receive the vacuum expectation values (vevs) $\langle \phi_1 \rangle$ and $\langle \phi_2 \rangle$, which must correspond to the global minimum of the potential. Therefore it is necessary that the extremum conditions are satisfied:

$$\left.\frac{\partial V}{\partial \phi_1}\right|_{\phi_1 = \langle \phi_1 \rangle} = 0, \quad \left.\frac{\partial V}{\partial \phi_2}\right|_{\phi_2 = \langle \phi_2 \rangle} = 0. \tag{3.16}$$

It should be stressed that the conditions (3.16) are only necessary, but not sufficient. The potential (3.15) might have several extrema or even several

minima. To ensure that the obtained values of $\langle\phi_1\rangle$ and $\langle\phi_2\rangle$ really correspond to the vacuum one must check that the vevs represent the minimum with the lowest value of the potential.

In the vacuum the global isospin symmetry can always be exploited to set the vevs into the following form:

$$\langle\phi_1\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v_1\end{pmatrix},\ \langle\phi_2\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}u\\ v_2 e^{-i\xi}\end{pmatrix}, \tag{3.17}$$

with v_1 and v_2 being real positive numbers, and $e^{-i\xi}$ some complex phase. Sometimes it is convenient to use polar representation for the v_1 and v_2:

$$v_1 = v\cos\beta,\ v_2 = v\sin\beta,\quad v = \sqrt{v_1^2 + v_2^2}. \tag{3.18}$$

The case $u \neq 0$ in (3.17) corresponds to the so-called charged vacuum [9]. While it can have some cosmological implications [8, 10], it cannot describe our world now, so we will only consider the case of neutral vacuum – the vacuum with $u = 0$.

The relative phase $e^{-i\xi}$ between vevs of the first and second doublet, it describes CP-violation in the 2HDM Higgs sector. The phenomenon of *spontaneous CP violation* (non-trivial value of ξ when the original lagrangian is explicitly CP-symmetric) was actually the historical reason to consider such a model.

In the neutral vacuum the extremum conditions (3.16) give a system of cubic equations in v_1 and v_2 (see e.g. [10]). The need to solve this system is a major obstacle for the analytic studies of the 2HDM phenomenology.

Supposing that we are, indeed, in the neutral vacuum, we decompose both fields around their vevs:

$$\phi_1 = \begin{pmatrix}\xi_1^+\\ \dfrac{v_1 + \chi_1 + i\xi_1}{\sqrt{2}}\end{pmatrix},\quad \phi_2 = \begin{pmatrix}\xi_2^+\\ \dfrac{v_2 + \chi_2 + i\xi_2}{\sqrt{2}}\end{pmatrix} e^{-i\xi}. \tag{3.19}$$

One can see that the combinations

$$G^\pm = \cos\beta\,\xi_1^\pm + \sin\beta\,\xi_2^\pm,\quad G^0 = \cos\beta\,\xi_1 + \sin\beta\,\xi_2, \tag{3.20}$$

are the massless Goldstone modes, while the combinations orthogonal to (3.20):

$$H^\pm = -\sin\beta\,\xi_1^\pm + \cos\beta\,\xi_2^\pm,\quad \chi_3 = -\sin\beta\,\xi_1 + \cos\beta\,\xi_2, \tag{3.21}$$

correspond to the charged Higgs boson and a neutral scalar with parity that is different from the parity of $\chi_{1,2}$. Linear combinations of neutral fields χ_i form the set of observable neutral Higgs particles h_1, h_2, h_3.

3.3.2 The Higgs Basis of the 2HDM

The model contains two fields with identical quantum numbers. Therefore, it can be described both in terms of fields ϕ_1, ϕ_2, used in Lagrangian (3.15) or in terms of fields ϕ_1', ϕ_2' obtained from ϕ_k by a global unitary *reparameterization (RPA)* transformation $\hat{\mathcal{F}}$ of the form:

$$\begin{pmatrix}\phi_1'\\ \phi_2'\end{pmatrix} = \hat{\mathcal{F}}\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\ \hat{\mathcal{F}} = e^{-i\rho_0}\begin{pmatrix}\cos\theta\, e^{i\rho/2} & \sin\theta\, e^{i(\tau-\rho/2)}\\ -\sin\theta\, e^{-i(\tau-\rho/2)} & \cos\theta\, e^{-i\rho/2}\end{pmatrix}. \quad (3.22)$$

This transformation induces another transformation $\lambda_i \to \lambda_i'$ of the parameters of the Lagrangian in such a way that the new Lagrangian in fields ϕ_i' have the same physical content. Different RPA-equivalent representations are referred to as "different RPA bases".

The unitary transformation (3.22) is parametrized by 4 angles θ, ρ, τ and ρ_0. The last parameter ρ_0 is irrelevant since it describes overall phase transformation of the fields without change of parameters of the potential – see e.g. [10].

It turns out that phenomenological analysis of the 2HDM gets much easier if we choose a particular RPA basis – the basis with $v_2 = 0$. Such a basis is called *the Higgs basis, or Georgi basis* (see e.g. [11], where it is used). It turns out that the analysis of the model is greatly simplified if one works in this basis.

The Higgs basis is obtained given vacuum v.e.v.s, using the reparametrization transformation (3.22) with $\theta = \beta$, $\rho - \tau = \xi$:

$$\hat{\mathcal{F}} = e^{-i\rho_0}\begin{pmatrix}\cos\beta\, e^{i\rho/2} & \sin\beta\, e^{i(\rho/2-\xi)}\\ -\sin\beta\, e^{-i(\rho/2-\xi)} & \cos\beta\, e^{-i\rho/2}\end{pmatrix}. \quad (3.23)$$

Free phase factor $e^{\pm i\rho/2}$ represent the *rephasing* freedom in the choice of the Higgs basis: the physical picture is independent on the choice of relative phase of ϕ_i.

The extremum conditions (3.16) in the Higgs basis are simplified to a pair of very simple relations (instead of a cubic system in a basis with $v_2 \neq 0$):

$$v^2\Lambda_1 = \tilde{m}_{11}^2, \quad v^2\Lambda_6 = \tilde{m}_{12}^2 \quad (3.24)$$

The decomposition (3.19) also gets simpler – the goldstone modes $G^{\pm}$, G^0, the charged scalar $H^{\pm}$ and the CP-odd scalar A become:

$$\phi_1 = \begin{pmatrix} G^+ \\ \dfrac{v+\eta_1+iG^0}{\sqrt{2}} \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} H^+ \\ \dfrac{\eta_2+iA}{\sqrt{2}} \end{pmatrix}. \tag{3.25}$$

Removing goldstone modes by employing the unitarity gauge (2.11), then substituting (3.25) into the potential (3.15) and examining the quadratic terms one obtains the mass of the charged higgs $M_{\pm}^2 = v^2\Lambda_3 - m_{22}^2$ and the mass matrix for neutral scalars in the (η_1, η_2, A) basis:

$$M_{ij} = v^2 \begin{pmatrix} \lambda_1 & Re\,\lambda_6 & -Im\,\lambda_6 \\ Re\,\lambda_6 & \frac{M_{\pm}^2}{2v^2} + \frac{\lambda_4+Re\lambda_5}{2} & -Im\,\lambda_5/2 \\ -Im\,\lambda_6 & -Im\,\Lambda_5/2 & \frac{M_{\pm}^2}{2v^2} + \frac{\lambda_4-Re\lambda_5}{2} \end{pmatrix}. \tag{3.26}$$

Diagonalization of this matrix results in three scalar eigenstates h_1, h_2 and h_3, which do not have a definite CP-parity. CP-conservation would mean that the mass matrix (3.26) is block-diagonal with $Im\,\lambda_6 = Im\,\Lambda_5 = 0$. The upper-left non-diagonal 2×2 block of the mass matrix is then diagonalized in the same way as it was done for (3.8) in electroweak singlet model. As a result one will obtain the usual two CP-even Higgs scalars h and H together with the CP-odd axial A.

3.3.3 Yukawa sector of the 2HDM

The most general form of the Yukawa sector of the 2HDM can be obtained by repeating the Yukawa construction (2.17) in the SM for each Higgs doublet:

$$\begin{aligned} -\mathcal{L}_Y = \Gamma_1^{ij} \begin{pmatrix} \bar{u}_L^i \\ \bar{d}_L^i \end{pmatrix} \phi_1\, d_R^j + \Delta_1^{ij} \begin{pmatrix} \bar{u}_L^i \\ \bar{d}_L^i \end{pmatrix} \tilde{\phi}_1\, u_R^j \\ + \Gamma_2^{ij} \begin{pmatrix} \bar{u}_L^i \\ \bar{d}_L^i \end{pmatrix} \phi_2\, d_R^j + \Delta_2^{ij} \begin{pmatrix} \bar{u}_L^i \\ \bar{d}_L^i \end{pmatrix} \tilde{\phi}_2\, u_R^j + h.c. \end{aligned} \tag{3.27}$$

In the most general case when $\Gamma_{1,2} \neq 0$ and $\Delta_{1,2} \neq 0$, the diagonalization procedure (2.18) cannot be simultaneously done for all four matrices. This implies an existence of generally large Flavor Changing Neutral Currents (FCNC). Observations tell us that FCNC processes are greatly suppressed, requiring unnatural tuning of the parameters of the 2HDM Yukawa sector.

To have a Natural Flavor Conservation (NFC) an extra symmetry of the 2HDM was proposed [12, 13]. The discrete Z_2 symmetry acts on the doublets in the following manner:

$$\begin{array}{c} \phi_1 \to -\phi_1 \\ \phi_2 \to \phi_2 \end{array} \quad \text{or} \quad \begin{array}{c} \phi_1 \to \phi_1 \\ \phi_2 \to -\phi_2 \end{array} \ . \tag{3.28}$$

This constraints the structure of (3.15) to two general alternatives:

$$\begin{aligned} \textbf{Model I:} \quad & \Gamma_2^{ij} = \Delta_1^{ij} = 0, \\ \textbf{Model II:} \quad & \Gamma_2^{ij} = \Delta_2^{ij} = 0. \end{aligned} \tag{3.29}$$

The Z_2 symmetry (3.28) also constrains the parameters of the potential: the $\lambda_{6,7}$ and m_{12}^2 terms in (3.15) violate that symmetry and must vanish for the potential to be Z_2 symmetric. (It must be noted, however, that the reparametrization basis (3.23) where the symmetry (3.28) is manifest, is not necessary coincident with the Higgs basis.)

The fully Z_2 symmetric 2HDM is too phenomenologicaly restrictive – usually the m_{12}^2 term is kept. Such quadratic terms are called called "soft symmetry-breaking" terms because they do not modify physics at high energies making the description stable under radiative corrections.

3.4 MSSM

Supersymmetric (SUSY) theories are based on introduction of an extra symmetry between fermions and bosons. That way one can complete the Grand Unification program, joining electroweak, strong and gravitational interactions into a single framework. Due to the spin-statistics theorem [14], bosons and fermions have integer and half-integer spins correspondingly. Therefore they belong to different representations of the Poincare group, which means that the generators of the supersymmetry should act "across" such representations.

For that purpose, the Lie algebra of Poincare group is extended to the so-called graded Lie algebra of "super-Poincare" group by introduction of extra anticommuting generators. It turns out that this is the only non-trivial way to extend Poincare group.

3.4.1 Minimal Supersymmetric extension of the SM

SUSY theories are arguably the most favorable by theorists among other classes of Beyond Standard Model extensions (or so it was for past several decades). Due to the introduced symmetry between bosons and fermions, the fermionic and bosonic degrees of freedom must come in pairs – every boson must have a corresponding suspersymmetric counterpart and vica-versa.

Standard Model, on the other hand, is very non-supersymmetric. It contains 96 (with massive neutrinos) fermionic and 28 bosonic degrees of freedom. The situation is worsened by the fact that even among these fields there are no candidates for a pair of superpartners in the SM – for every fermion there is no bosons with similar quantum numbers. This means that any SUSY extension of the SM must provide a super partner for *every* field in the SM. The obvious contradiction with observation is then resolved by introduction of the Λ_{SUSY} – the SUSY breaking scale, below which the symmetry is broken and only some of the particles can be observed.

The boson-fermion mirroring is very attractive from theoretical point of view even if it happens at Λ_{SUSY}, because this resolves the hierarchy (see Chap. 3.1.3) problem: every bosonic loop on Fig. 3.2 would have the corresponding contribution with the boson's supersymmetric counterpart, and the fermion loop will get the similar bosonic contribution. Bosonic and fermionic loops have opposite signs, so these loop pairs would cancel each other at the SUSY scale.

The number $\mathcal{N}$ of the introduced supergenerators is generally unconstrained, leaving a number of theoretical possibilities at high scales. However at the TeV scale SUSY theory with $\mathcal{N} > 1$ cannot be chiral, which will spoil the anomaly cancellation mechanics (2.21) of the SM. In general we can consider a model with $\mathcal{N} > 1$ at some higher scales. Nevertheless for the low energy effective SUSY extention of the Standard Model we must have $\mathcal{N} = 1$.

3.4.2 Higgs sector of the MSSM

The superpartners of gauge bosons are Majorana fermions – they are left-right symmetric and do not contribute to chiral anomalies. However, this is not the case for the superpartner of the Higgs scalar. The Higgs superpartner is called "higgsino", that fermion also contributes to fermionic loops on Figure 2.2; and must be canceled in order to avoid inconsistencies in the the-

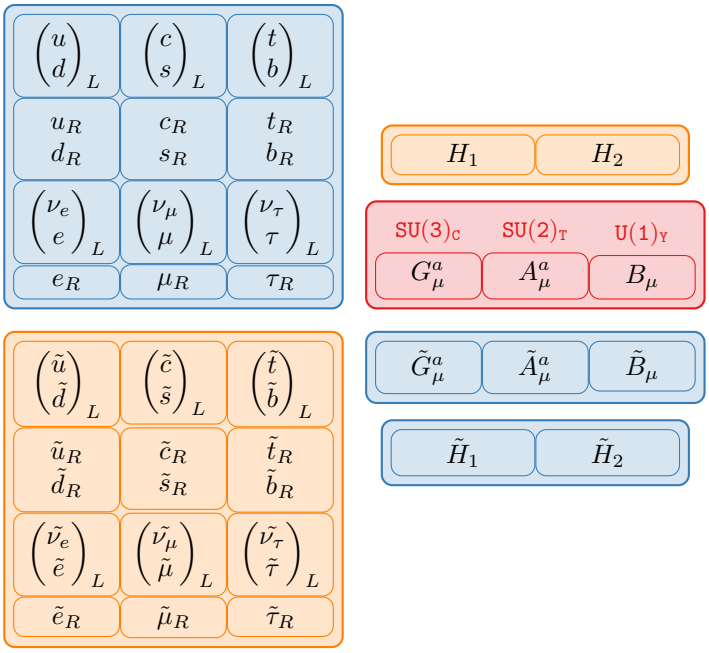


Figure 3.3: Particle content of the MSSM.

ory. For this reason MSSM introduces two oppositely charged higgsinos and, therefore, two Higgs doublets with opposite hypercharges:

$$H_1 = \begin{pmatrix} H_1^0 \\ H_2^+ \end{pmatrix}, \quad H_2 = \begin{pmatrix} H_1^- \\ H_2^0 \end{pmatrix}. \tag{3.30}$$

The symmetry considerations require the Higgs-fermion Yukawa interactions of MSSM to have the Model II structure. The quartic parameters λ_i of the MSSM higgs potential are not the free parameters of the model as in (3.15) (or even in (2.8)), but are fixed by gauge interaction couplings. The Higgs potential of the MSSM reads:

$$\begin{aligned} V = &\frac{g^3 + g'^2}{8}\left(H_1^\dagger H_1 - H_2^\dagger H_2\right)^2 + \frac{g^2}{2}\left(H_1^+ H_2\right)^2 \\ &+ m_1^2\, H_1^\dagger H_1 + m_2^2\, H_2^\dagger H_2 - 2\, m_3^2\, Re\, H_1 H_2\,, \end{aligned} \tag{3.31}$$

Therefore, in the MSSM Higgs sector is the special case of the 2HDM with fixed couplings. The quadratic parameters m_i^2 in (3.31) are also fixed at Λ_{SUSY},

but their values are such that the potential has only a trivial minimum. The radiative symmetry breaking is expected to occurs when one calculates the RG flow of the parameters m_i^2 down to electroweak scale. Still the model will not break the CP-symmetry in contrast to more general 2HDM.

3.4.3 MSSM Benchmark scenarios

The general MSSM contains 126 free parameters – study of such an enormous parameter space is unfeasible. The dimension of parameter space can be substantially reduced by various symmetry requirements and needs to conform with experimental observations both from collider experiments and from cosmology.

The resulting parameter space is much smaller ($\sim 5 \div 10$ parameters), but still is too large to be studied right away. It is still necessary to select a number of representative "benchmark" points highlighting various phenomenological possibilities of the model. It is preferable that such benchmark points would provide easily observable signatures at colliders, so that the regions of parameter space could be falsified.

This is especially interesting in light of electroweak physics at colliders. MSSM might provide a light Higgs boson, which might be easier to detect than the Higgs boson in the SM. The MSSM Higgs sector is fully determined by the three quadratic parameters m_i^2, which are usually represented in terms of move conventional v^2, $\tan\beta$ and M_A. Minimization of the potential (3.31) gives for the masses of CP-even Higgses:

$$m_{H,h}^2 = \frac{1}{2}\left(m_A^2 + m_Z^2 \pm \sqrt{(m_A^2 + m_Z^2)^2 - 4m_Z^2 m_A^2 \cos^2 2\beta}\right). \tag{3.32}$$

The lightest Higgs mass, therefore, satisfies $m_h \leq m_Z$ at tree level. Such a light Higgs is long excluded by collider experiments. Still, it is noticeable that The m_h mass, is also influenced by radiative corrections from $\tilde{t}$ (and from $\tilde{b}$ at large $\tan\beta$); nevertheless, MSSM favors quite light Higgs masses and, in view of LEP and Tevatron bounds, the so-called m_h^{max} benchmark scenario was suggested [15, 16] to be used for LHC searches.

The m_h^{max} benchmark targets at maximizing the supersymmetric Higgs masses, while keeping parameters of the benchmark point transparent and with their values and brief description follow:

- $M_{SUSY} = 1$ **TeV** is the value, fixing masses of the squarks. As it turns out the upper bound on m_h is fairly independent of the relation between squark masses. Being sensitive only to the heaviest one.
- $X_t = 2M_{SUSY}$ is the stop mixing parameter – the off-diagonal values of the stop mass matrix.
- $\mu = 200$ **GeV** the higgsino mass.
- $M_{\tilde{g}} = 800$ **GeV** the gluino mass.
- $M_2 = 200$ **GeV** the SU(2)-gaugino mass parameter.
- $A_b = A_t$ are the stop and sbottom trilinear couplings, respectively.

Bibliography

[1] Edward W. Kolb and Michael S. Turner. "The Early Universe". *Front.Phys.* 69 (1990), pp. 1–547.

[2] P.A.R. Ade et al. "Planck 2013 results. XVI. Cosmological parameters" (2013). arXiv:1303.5076 [astro-ph.CO].

[3] Jerome Martin. "Everything You Always Wanted To Know About The Cosmological Constant Problem (But Were Afraid To Ask)". *Comptes Rendus Physique* 13 (2012), pp. 566–665. arXiv:1205.3365 [astro-ph.CO].

[4] Joan Elias-Miro et al. "Higgs mass implications on the stability of the electroweak vacuum". *Phys.Lett.* B709 (2012), pp. 222–228. arXiv:1112.3022 [hep-ph].

[5] Robert Schabinger and James D. Wells. "A Minimal spontaneously broken hidden sector and its impact on Higgs boson physics at the large hadron collider". *Phys.Rev.* D72 (2005), p. 093007. arXiv:hep-ph/0509209 [hep-ph].

[6] Vernon Barger, Danny Marfatia, and Andrea Peterson. "LHC and Dark Matter Signals of Z' Bosons". *Phys.Rev.* D87 (2013), p. 015026. arXiv:1206.6649 [hep-ph].

[7] T.D. Lee. "A Theory of Spontaneous T Violation". *Phys.Rev.* D8 (1973), pp. 1226–1239.

[8] I.F. Ginzburg, I.P. Ivanov, and K.A. Kanishev. "The Evolution of vacuum states and phase transitions in 2HDM during cooling of Universe". *Phys.Rev.* D81 (2010), p. 085031. arXiv:0911.2383 [hep-ph].

[9] A. Barroso et al. "Stability of the normal vacuum in multi-Higgs-doublet models". *Phys.Rev.* D74 (2006), p. 085016. arXiv:hep-ph/0608282 [hep-ph].

[10] I.F. Ginzburg and K.A. Kanishev. "Different vacua in 2HDM". *Phys.Rev.* D76 (2007), p. 095013. arXiv:0704.3664 [hep-ph].

[11] C. C. Nishi. "Physical parameters and basis transformations in the two-Higgs-doublet model". *Phys. Rev. D* 77 (5 2008), p. 055009.

[12] Sheldon L. Glashow and Steven Weinberg. "Natural conservation laws for neutral currents". *Phys. Rev. D* 15 (7 1977), pp. 1958–1965.

[13] Emmanuel A. Paschos. "Diagonal neutral currents". *Phys. Rev. D* 15 (7 1977), pp. 1966–1972.

[14] W. Pauli. "The Connection Between Spin and Statistics". *Phys. Rev.* 58 (8 1940), pp. 716–722.

[15] Marcela S. Carena et al. "Suggestions for benchmark scenarios for MSSM Higgs boson searches at hadron colliders". *Eur.Phys.J.* C26 (2003), pp. 601–607. arXiv:hep-ph/0202167 [hep-ph].

[16] Marcela S. Carena et al. "MSSM Higgs boson searches at the Tevatron and the LHC: Impact of different benchmark scenarios". *Eur.Phys.J.* C45 (2006), pp. 797–814. arXiv:hep-ph/0511023 [hep-ph].

Chapter 4

The Large Hadron Collider

The Large Hadron Collider (LHC) is the world lagrest particle accelerator, capable of producing the highest energies achieved by humankind so far. It was constructed by the European Organization for Nuclear Research (CERN) – an international organization, that joins more than 600 institutes and universities around the world.

The LHC is set up [1] in a circular tunnel formerly occupied by the Large Electron Positron (LEP) collider. The tunnel circumfence is about 27km and it is located about 100m underground. It now contains more than 1600 superconducting magnets for bending and focusing the circulating beams of charged particles.

In the proton mode, two counterpropagating proton beams are circulating it the tunnel. The design goal for the LHC is to achieve energies of 7TeV per nucleon, thus reaching $\sqrt{s} = 14\,TeV$ center-of-mass collision energy.

Such energies are required by the need of studying the TeV electroweak scale – protons are composite particles and the center-of-mass collison energy of individual partons is smaller than the energy of the protons.

4.1 Experiments at the LHC

In total, there are seven different detectors, performing various experiments at the LHC. Three of them, TOTEM, MoEDAL and LHCf, are quite small and are designed for very specialized research. Figure 4.1 shows the the LHC layout [2] and how four large experiments are located on the different straight

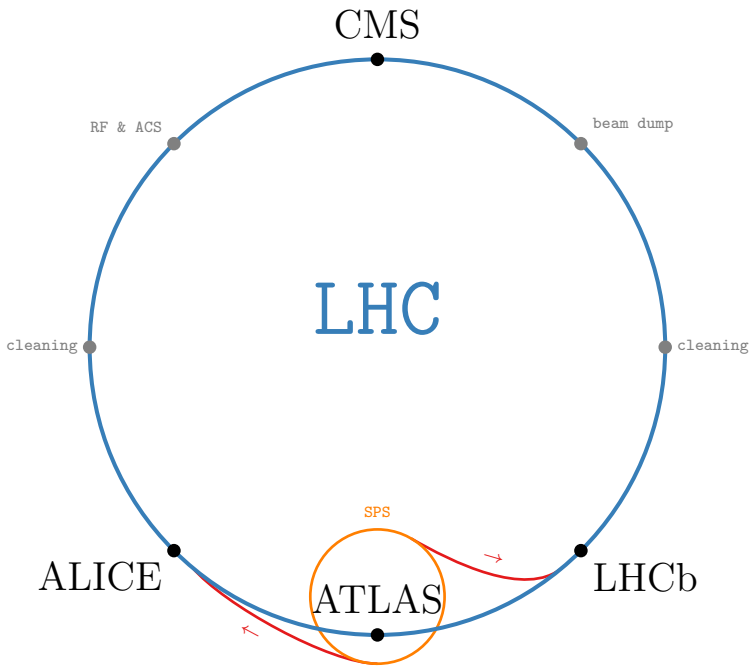


Figure 4.1: The LHC layout

sections of the LHC ring.

4.1.1 A Large Ion Collider Experiment (ALICE)

Created to study QCD phenomena in Ion-Ion collisions at center-of-mass energies of about $\sqrt{s} = 2.76$ TeV per nucleon. The corresponding temperatures and energy densities are deemed to be the highest in any of the physics experiments so far. At such temperatures the matter is expected to be in the state called the Quark-Gluon Plasma (QGP), in witch quarks and gluons are deconfined.

The detector [3] is specifically optimized to precisely register tens of thousands of particles, produced in such collisions. Very high granularity, immense data aqusition throughput and strong emphasis on particle identification methods are the distinguishing features of the detector.

4.1.2 A Toroidal LHC Apparatus (ATLAS)

Largest in size detector at LHC, ATLAS is one of the two general-purpose detectors. It [4] follows the standard scheme of a detector design: inner tracker, surrounded by electromagentic and hadronic calorimeters (ECAL and HCAL), which are in turn surrounded by muon system.

ATLAS invested in better e/γ reconstruction efficiency and jet resolution. It is recognized for its liquid argon ECAL, "standalone" muon subsystem and two-parts magnetic system with huge toroidal magnet coils surrounding the detector.

4.1.3 Compact Muon Solenid (CMS)

The second general-purpose detector at the LHC that also follows the standard tracker-ECAL-HCAL-muon system scheme.

The detector invested in better e/γ and μ momentum resolution, introducing higher magnetic field and all-silicon inner tracker. The CMS is distinguished for its two-level (instead of convetional 3 levels) trigger architecture and the world-largest magnetic field density in a compact solenoid that provides 4T magnetic field.

4.1.4 Large Hadron Collider beauty experiment (LHCb)

The experiment is optimized to study the properties of B-mesons – hadrons that contain the b-quark. Unique properties of the b-quark help in precise studies of flavor physics and CP-violation.

The detector [5] is recognized for its unusual forward-conic layout, dipolar bending magnet and use of of Ring Image CHerenkov (RICH) particle identification behind the tracker.

4.2 The LHC luminocity and PileUp

The two most important quantitative characteristics of a particle collider are its center-of-mass collision energy and its luminocity. Luminocity, (or, more precisely, instantaneous luminocity) $\mathcal{L}(t)$ is defined as an integral characteristic of an experiment that relates the cross-section σ of a given process and

an average rate at which events for this process are produced:

$$N = \sigma \cdot \int \mathcal{L}(t)\, dt. \tag{4.1}$$

The typical electroweak process we are interesed in, is the producion of a pair of massive vector bosons: WW or ZZ. The corresponding cross-section can be estimated to be around $\sigma \sim \alpha_W^2/s$, which is about 500 fb at per-parton energies $\sqrt{s} \sim 1$ TeV. That value is further reduced by the requirement of the produced bosons to decay into pairs of leptons – the detection sensitivity is greatly reduced in case of hadronic decays due to huge background of QCD produced jets. The leptonic branching ratio of the W and Z bosons is roughly $BR_{lept} \sim 1/9$ so the final estimate for the typical cross-sections we are intersted in is $\sigma \cdot BR_{lept}^2 \sim 5 \div 10$ fb.

To obtain sufficient statistics for such processes at least $N = 1000$ of yearly events are required. This, using (4.1), puts a requirement on the integrated and instantaneous luminocity of the experiment:

$$\int_{year} \mathcal{L}(t)\, dt = 100\text{fb}^{-1}, \quad \mathcal{L} = 10\frac{1}{\text{nb} \cdot \text{sec}}, \tag{4.2}$$

assuming that the running time takes about $1/3$ of each year.

4.2.1 Pileup

At LHC every beam is subdivided in a sequence of bunches. If each bunch contains n particles and bunches are crossing head-to-head with frequency f, then the luminocty of such an operation is given by:

$$\mathcal{L} = f\frac{n^2}{4\pi\sigma_x\sigma_y}, \tag{4.3}$$

where σ_x and σ_y describe the transverse distribution of the particles in the bunches. The expression (4.3) makes it clear that in order to increase the luminosity one could increase the number of particles in bunch, improve the focusing of the beams or increase frequency of bunch-crossings.

The frequency f is bounded by the maximal speed of the detector that would still allow temporal resolution of consecutive bunches. For example one can estimate (see e.g.[6]) that the charge collection time of a silicon solid

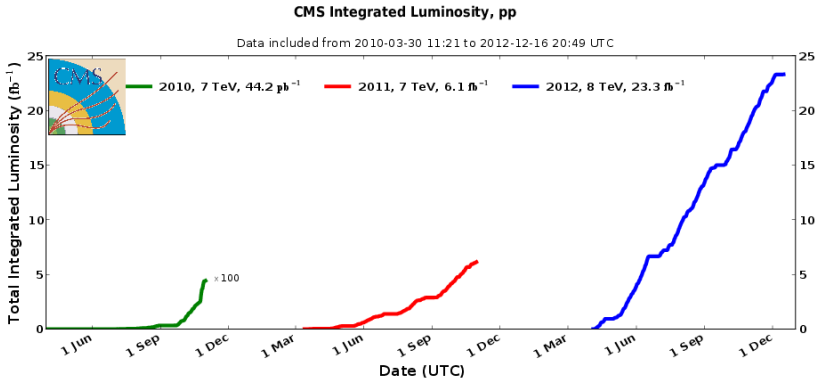


Figure 4.2: Delivered LHC luminocity through years 2010–2012.

state detector is about $\sim$ 20 ns. Therefore it will be very inefficient to adopt bunch crossings happening more rapidly than every 20 ns – in that case the detector will just integrate over multiple bunch crossings.

The bunch spacing at LHC is targeted to be 25 ns between consequent bunches. The total inelastic pp cross-section [7] at LHC energies is about $\sigma_{tot} \sim 100$ mb: this means that at the target instantaneous luminocity (4.2) one should expect one inelastic interaction per nanosecond, or ~ 25 inelastic interactions per bunch crossing. This interactions are unavoidable if one targets ath the LHC design goals. They constitute the so-called *"PileUp"*, which is an important experimental concern for the analyses at the LHC.

More detaied descritpion of the PileUp treatment in various reconstruction stages at the CMS detector will be discussed further in the Chapter 6

4.2.2 LHC operation during 2011-2012

The timeline for the LHC operation expects a gradual approach to the nominal energy and luminocity. The start of the experiment was planned on September of 2008 – the proton beams started to successfully circulate the main ring, but after 9 days the failure of electric connection in one of the magnets resulted in loss of the superconducting state. Subsequent rapid heating of the magnet lead to explosion of cryogenic liquid helium, damaging surrounding magnets and delaying the start of normal operation by an extra year.

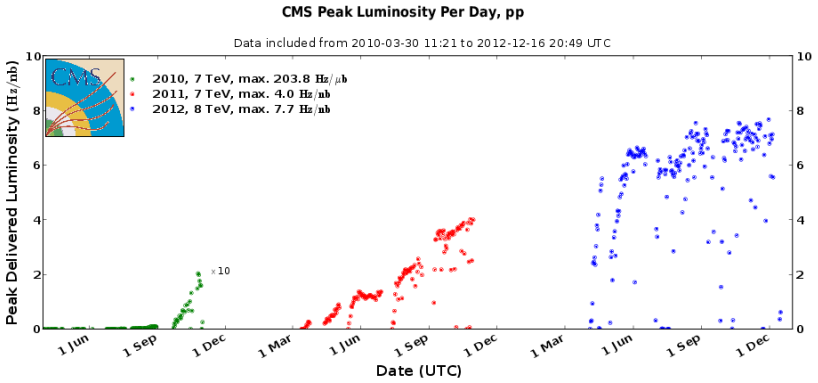


Figure 4.3: Peak intantaneous LHC luminocity through years 2010–2012.

First collisions with stable proton beams with center-of-mass energy $\sqrt{s} =$ 7 TeV have started on March 13, 2011. The collisons continued until November of 2011 and provided ATLAS and CMS experiments with about 5 fb^{-1} of data, see Figure 4.2. The second proton-proton run continued from May to November of 2012. The center-of-mass energy in 2012 was $\sqrt{s} = 8$ TeV. As it is shown on Figure 4.2, the total recorded data at the CMS detector is 21 $fb^{-}1$.

Throughout the years 2010 – 2011 the instantaneous luminocity was not constant, but slowly increasing for each operation period (see Figure 4.3). The gradual increase of the instantaneous luminocity and buch-crossing frequency is very useful for calibration and comissioning of the detectors and also for measurement of the SM process before the design LHC values are reached: currently the LHC is in its first long shutdown (LS1) phase. New

	2011	2012	Design
Beam Energy	3.5 TeV	4 TeV	7 TeV
Delivered Luminocity	6.1 fb^{-1}	23.3 fb^{-1}	
Bunch spacing	75 ns,50 ns	50 ns	25 ns

collisions at eneries $\sqrt{s} = 14$ TeV and separation of 25ns are expected to be started in the year 2015. During the two years of the LS1, series of renovations in the LHC infrastructure are being performed, incuding improvements in the accelerator itself and numerous upgrades of the detectors.

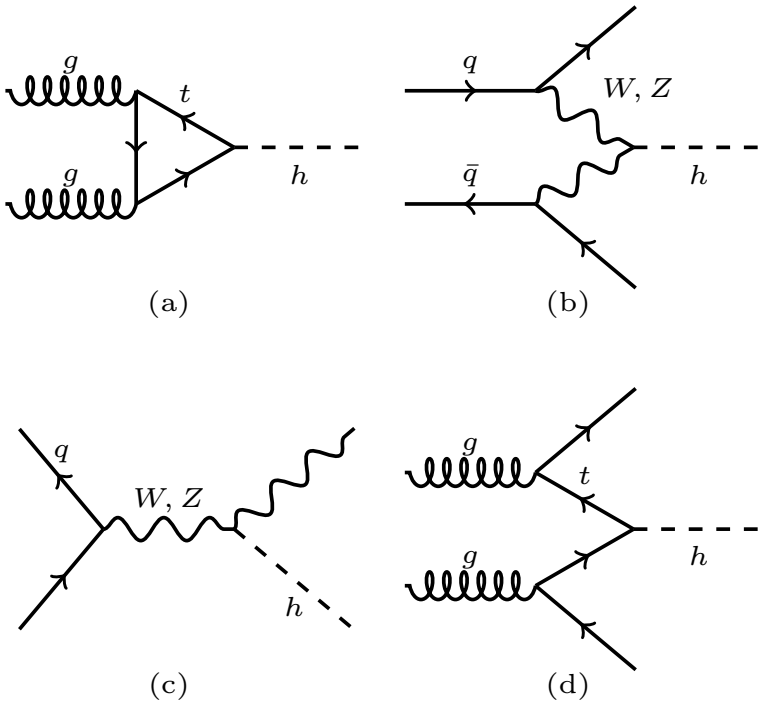


Figure 4.4: The four production modes of the SM Higgs at the LHC.

4.3 Higgs boson production at the LHC

Unitil the summer of 2012 the Higgs particle was considered a "missing piece" of the Standard Model. Discovery on the Higgs boson was one of the most important goals for the LHC machine.

Figure 4.4 shows the four main processes, contributing to the Higgs production at the LHC: the gluon fusion process (a), the vector boson fusion (VBF) process (b), the vector boson associated production (c) and the t-quark associated production (d). The corresponding cross-sections depend both on the mass of the Higgs boson m_h and on the center-of-mass collision energy $\sqrt{s}$. Typically, the cross-section decreases with m_h and increases with $\sqrt{s}$.

Figure 4.3 shows the cross-sections for the four production processes for $\sqrt{s} = 8$ TeV. The numbers are the result of the combined effort of the Higgs Cross-Section working group [8–10].

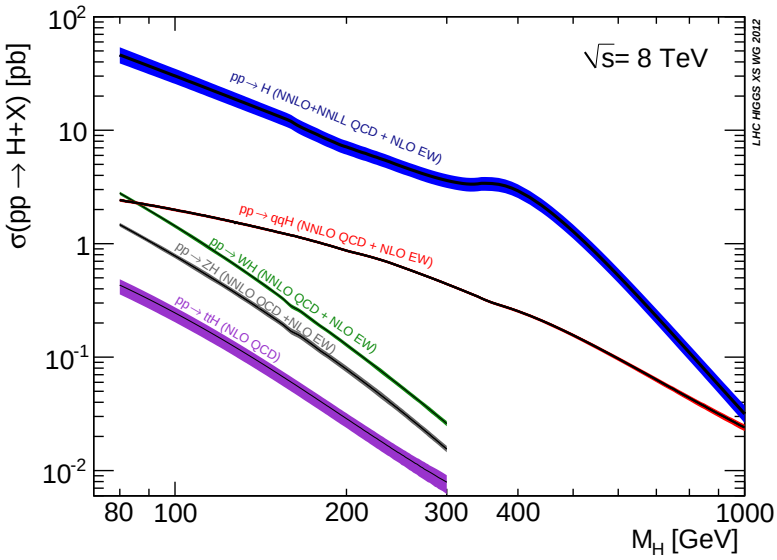


Figure 4.5: The production cross-sections for the four modes, depending on the HIggs mass at $\sqrt{s} = 8$ TeV.

Gluon fusion provides the largest production cross-section for the Higgs particle in the whole mass range. Main contribution to the cross section comes from the t-quark loop because of the strong Higgs-top Yukawa coupling. The process, therefore can be sensitive to values of Yukawa couplings and to extra fermions that might contribute to the loop (like fourth generation fermions).

The gluon fusion cross-section is very sensitive to radiative corrections – the factor of two increase is observed changing from leading order (LO) to next to leading order (NLO) calculation. The NNLO corrections are calculated in the large-m_t limit, adding extra 25% and, finally, soft gluon contributions are resummed up to NNLL, amounting for an extra 5% increase in the cross-section.

VBF production cross-section is an order of magintude smaller but searches can benefit from more easily recongnizable signature – two extra jets from hadronization of the associated quarks have large invariant mass and lie in more forward rapidity regions. VBF mechanism provides a proof of trilinear couplings at tree level, which is essential for determinig the nature of the Higgs sector of the SM.

The cross-sections for the two extra production mechanisms: ***Higgstrahlung*** and ***t*-quark associated** production are even smaller, but have a very specific final state signatures, that might be more easily recognized in the searches. Cross-sections of this processes can also be enanced in some BSM extensions, offering a potential for exclusions of these models before the sensitivities for the Standard Model Higgs productions are reached.

Bibliography

[1] Oliver S. Bruning et al. "LHC Design Report. 1. The LHC Main Ring" (2004). Ed. by Oliver S. Bruning.

[2] Lyndon Evans and Philip Bryant. "LHC Machine". *Journal of Instrumentation* 3.08 (2008), S08001.

[3] The ALICE Collaboration. "The ALICE experiment at the CERN LHC". *Journal of Instrumentation* 3.08 (2008), S08002.

[4] The ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". *Journal of Instrumentation* 3.08 (2008), S08003.

[5] The LHCb Collaboration. "The LHCb Detector at the LHC". *Journal of Instrumentation* 3.08 (2008), S08005.

[6] Dan Green. "At the leading edge: The ATLAS and CMS LHC experiments" (2010). Ed. by Dan Green.

[7] CMS Collaboration. "Measurement of the inelastic *pp* cross section at sqrts = 7 TeV" (2011).

[8] LHC Higgs Cross Section Working Group et al. "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables". *CERN-2011-002* (CERN, Geneva, 2011). arXiv:1101.0593 [hep-ph].

[9] LHC Higgs Cross Section Working Group et al. "Handbook of LHC Higgs Cross Sections: 2. Differential Distributions". *CERN-2012-002* (CERN, Geneva, 2012). arXiv:1201.3084 [hep-ph].

[10] LHC Higgs Cross Section Working Group et al. "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties". *CERN-2013-004* (CERN, Geneva, 2013). arXiv:1307.1347 [hep-ph].

Chapter 5

CMS detector

The Compact Muon solenoid (CSM) detector is one of the two general purpose detectors at the LHC, desingend to study New Physics at the electroweak scale. It follows the "generic" detector layout [1], arranged to provide high granularity and hermetic coverage of ± 5 units in pseudorapidity.

The coordiante system used at the LHC is the right-handed coordinate system. The origin is positioned at the nominal collision point, x-axis is directed to the center of the LHC ring, the y-axis points up, and the z-axis goes in the anticlockwise direction of the LHC beamline. More convenient cylindrical coordinate system (r, ϕ, η) is routinely used. Pseudorapidity η is a coordinate, that describes the polar angle with respect to the beam axis and approximates rapidity in the ultra-relativistic limit:

$$\eta = -\log\left(\tan\theta/2\right) \quad \sim y \text{ for } E \gg m. \tag{5.1}$$

5.1 CMS subsystems

In order to measure all the particles produced in the high energy collisions at the LHC, the detector should be made of several cooperating subsystems: a *tracker* stays as close to the the beam as possible, measuring tracks of charged particles and determining positions of primary and secondary vertices. The tracker is followed by two calorimeters: electromagentic (ECAL) and hadronic (HCAL). Both calorimeters capture hadrons and charged particles, measuring their positions and energies. The farthest subsystem is the muon tracking system – high transverse momentum muons penetrate both calorimeters and

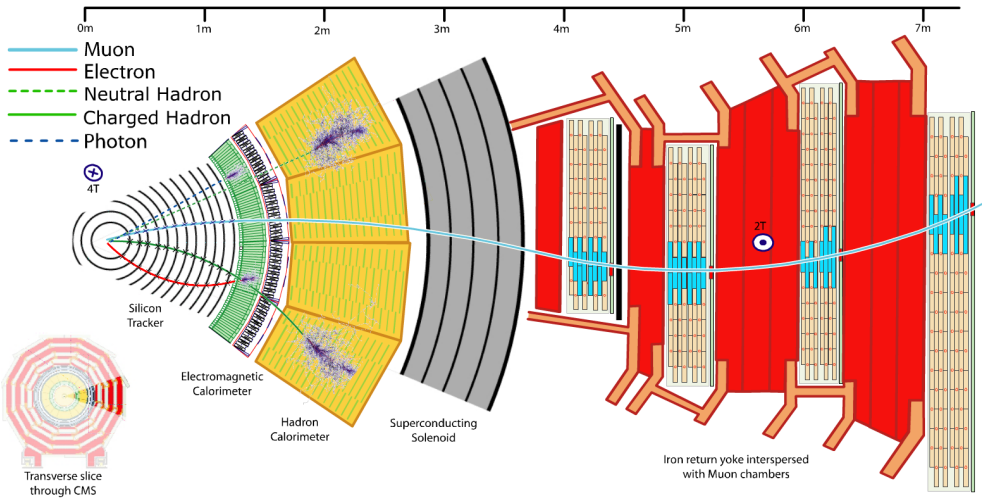


Figure 5.1: Slice ($r \times \phi$) of the CMS detector, showing all the detector's subsystems.

get measured in the muon chambers. Finally, a magnetic system must be incorporated in the detector design, providing bending magnetig field for trackers.

Figure. 5.1 shows the transverse slice of the CMS detector. Locations and radial spans of all the components mentioned above are shown. Example trajectories of various SM particles and their signatures in the CMS components are schematically drawn on the figure.

The longitudinal structure of a detector is mainly determined by radiation field. Radiation gets stronger for more collinear processes – that is why more sensitive detecting hardware is limited to the so-called "central" region of pseudorapidities $|\eta| < 2.5$. Both tracking and precision calorimetry are constrained to central rapisities by the harsh radiation in more forward regions. As a result, geometry of the generic detector is usually cylindrical: a barrel, covering $|\eta| < 1.5$ and two endcaps between $1.5 < |\eta| < 2.5$. Even more forward regions, down to $|\eta| < 5$ are covered by two forward hadron calorimeters that are designed to be more radiation-resistant.

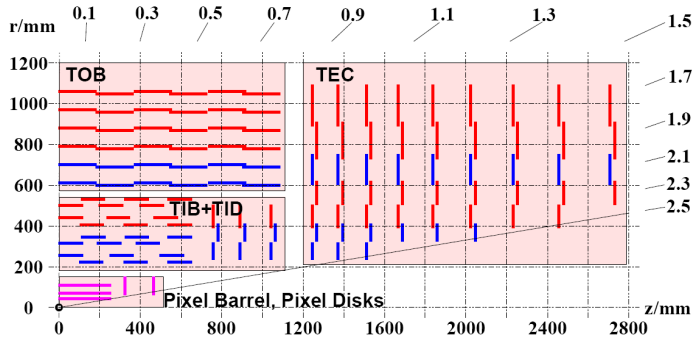


Figure 5.2: Slice ($r \times z$) of one quadrant of inner tracking system of the CMS detector.

5.1.1 CMS Tracker

Charged particle tracks reconstruction and determination of vertex positions is the basic step in the full event reconstruction process. Efficient tracking and high resolutions close to the interaction point are also especially important for b-tagging algoritms, discussed in Chapter. 7.

High resolution/granularity, fast response times and resistance to radiation damage (and, also, the cost constraints) was all contributing to the final design of the CMS tracker [2]. Different detector technologies was employed at different radial ranges from the beam, each providing the best balance between the mentioned requirements: Pixel Detector in the high-density region below $r \sim 20$ cm, Silicon Strip Tracker (SST) covers both medium-density region betwen $20 \lesssim r \lesssim 50$ cm. The low-density region lies in between $50 \lesssim r \lesssim 120$ cm where Larger-Pitch Silicon Strips are used.

Figure 5.2 shows the layout of the tracker subsystems, demonstrating the geometry configuration for each of the subtrackers – each contains a central barrel and a pair of endcaps, thus covering $|\eta| < 2.5$ pseudorapidity range.

Pixel Detector has three layers in the barrel and two bi-layer endcaps. The barrel pixel layers are situated at distances of $r = 4.4$ cm, $r = 7.3$ cm and $r = 10.2$ cm from the beam axis (bounded by radiation resistance from below and by cost from above). Four endcap disks at $|z| = 34.5$ cm and $z = 46.5$ cm cover more forward pseudorapidity regions. The size of each pixes is chosen to be 100×150 μm^2, providing the resolution of about 10 μm in transverse plane and 15 μm along the z-axis. The overall occupancy of

about 1% is achieved.

Silicon Strip Tracker completes the inner tracker of the detector. SST is the Largest silicon tracker ever built covering a total 200 m^2 of silicon sensors. Silicon microstrips of minimum dimension 10 cm μm are used at ranges between 20 cm and 55 cm, building up the Tracker Inner Barrel (TIB) and Tracker Iner Disks (TID) – see Figure. 5.2. Larger high-pitch silicon strips with maximum size about 25 cm μm are used on outer range of the tracker, in the Tracker Outer Barrel (TOB) and Tracker EndCaps (TEC), also shown on Figure. 5.2.

The CMS tracker is made entirely of silicon. That makes it world largest silicon detector and the first detector in particle physics to use silicon for outer tracker.

5.1.2 ECAL

Electromagnetic Calorimeter [3] at the CMS starts right after the inner tracker. The ECAL at CMS is a homogeneous calorimeter – the whole volume of the sumbdetector is sensitive to passing particles and contributes to the recorded signal.

The radial size of the calorimetry is determined by the characteristic interaction distance in the capturing medium. The medium used at the ECAL of the CMS detector is another novelty for detectors at the hadron colliders – it is made of lead thungstate ($PbWO_4$) crystals. ECAL's barrel and both endcaps agregate in total 75000 crystals. Each about 30cm long 2.5cm round – which roughly corresponds to electron's moliere radius and interaction length.

In order to aid neutral pion and electron identification and to imporve the position resolution for charged paritcles in the endcaps: $1.479 < |\eta| < 3.0$ A *preshower* device is installed before the ECAL crystals at pseudorapidity ranges $1.65 < |\eta| < 2.6$. The preshower contains two layers – lead radiators that start the showers from incoming particles and then, immediately after them, strip sensors that measure the resulting shower energy and shape.

5.1.3 HCAL

The CMS Hadron Calorimeter [4] supplements the Electromagnetic Calorimeter, building up complete calorimetry system for jets and Missing Transverse Energy.

In the "big picture" of the CMS concept, several choices was made in order to firstly maximize the physics performance but also to complement the ATLAS – the second general-purpose detector at the LHC [5]. From that point of view CMS is designed with emphasis on ECAL and muon resolution with a very strong manget. This strongly constrains the design on the HCAL, since it has to be placed partly outside and partly inside the solenoid. That obviously reduces the efficiency of the calorimetry, loosing in that respect to the ATLAS detector.

HCAL design is forced to have high density, to be nonmagnetic, and to operate in the 4T magnetic field. Layers of brass are used at the HCAL as an absorber. They are interlayed whith plastic scintillator as an active material. The scintillation light is gathered by fibers and directed to photodiodes that are also located inside the magnet – any alternative location would result in inacceptable light loss in transfer fiber optics. The choice of readout photodetectors was therefore constrained by the ability to operate in high magnetic field. Phototubes are unsuitable in such conditions, since they loose their gain due to inability to focus the electrons. The hybrid photodiodes was chosen instead to read out the singlas.

The outer hadron calorimeter (HO) utilizes the CMS magnet coil/cryostat and the steel of the magnet return yoke as its absorber. Array of scintillators and readout photodetectors, located just outside the magnet. The inclusion of HO layers improves shower containment in the barrel region $|\eta| < 1.26$. HO uses the same active material and readout system as the barrel and endcaps.

To have a good missing transverse energy determination it is crucial to capture all the hadrons produced in the pseudorapidity range $|\eta| < 5$. The separate set of two forward hadron calorimeters (HF) is utilized for the forward region $3.0 < |\eta| < 5.0$. Forward calorimenters are made of quartz fiber that generate Cherenkov light, while embeeded in iron absorber – the HF is located outside the magnetic field volume, allowing for the use of magnetic materials. To that end, Photomutliplier Tubes are also used to read out the signal.

Figure. 5.3 (sourced [6]) concludes our discussion of the calorimetry at CMS, showing comparative efficiencies of both ECAL and HCAL at various detectors at the CMS (see Chapter. 4.1). One can see the superior resolution of homogeneous electromagnetic calorimeter at CMS (the ATLAS's ECAL is sampling, using liquid argon as its medium). Also it is clear from Figure. 5.3

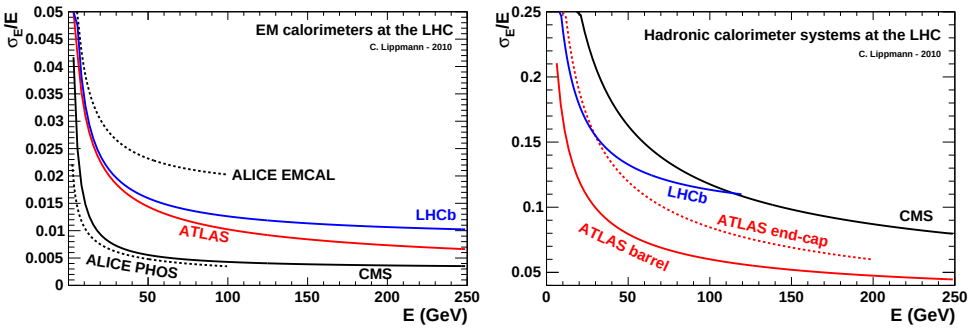


Figure 5.3: Comparison of energy resolutions of electromagnetic (left) and hadronic (right) calorimeters at the LHC detectors.

how HCAL resolution is degenerated as a price for limited volume, required by the magnet.

5.1.4 Muon system

There are only two types of particles that are likely to penetrate volumes of the tracker, both calorimeters and solenoid – neutrinos and muons. While neutrinos are completely escaping the dector, leaving only indirect evidence of their presense in from of the missing transverse momentum. Muons still can be detected and used to improve the gathered data.

Muon system the CMS detector is the outer subsystem of the detector, serving two main purposes. Firstly, it complements the inner tracker, improving the muon momentum measurement for muons with high transverse momentum. Second important function of the muon susbsystem is its use in the fast decision making at L1 trigger for events with muon in final state (see next Section).

Three types of particle detectors for muon identification are employed: drift tubes chambers (DT) for central region $|\eta| < 1.2$, cathode strip chambers (CSC) for endcaps $0.9 < |\eta| < 2.4$ and resistive plate chambers (RPC) both for barrel and endcaps – DT and CSC types have an advantage in spatial resolution. While the RPCs, on the other hand, win with respect to their timing and used association of registerd muons to bunch crossing and for triggering.

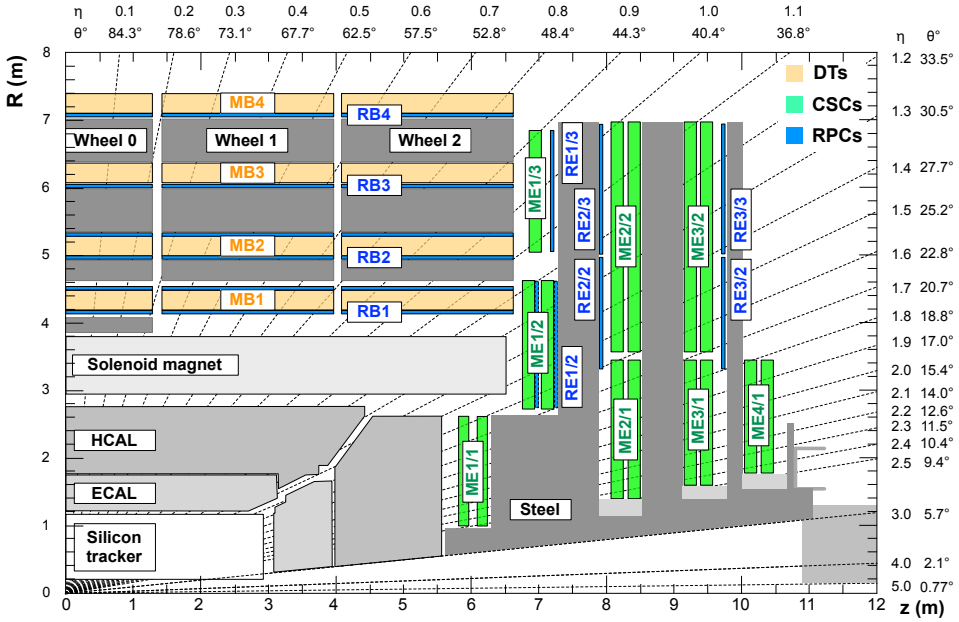


Figure 5.4: Slice ($r \times z$) of one quadrant of the muon subsystem, tracker, both calorimeters and the solenoid are also shown.

5.1.5 Magnet

High resolutions both for inner tracker and for muon system are relying on the strong bending power of the magnetic field, which effectively increases the detector's granularity. Such a srong magentic field is provided by the Compact Solenoid of the CMS detector. The magnet [7] is designed to achieve 4T magnetic field inside the solenoid, which surrounds the tracker, the ECAL and the larger part of the HCAL.

The solenoid's diameter is 6m and it is 13.5m long. The so-called *return yoke* – the 12-sided structure made of iron "wheels" – is used to return the magnetic flux by containg and guiding the field. The yoke reaches 14m in diameter, it is split in three layers, traversing the CMS's muon system, also helping to block all the remaining non-muon particles that wasn't captured by calorimetry.

The magnet is the largest superconducting magnet ever built, the operating magnetic field for 2011-2012 runs was scaled down to 3.8T to minimize

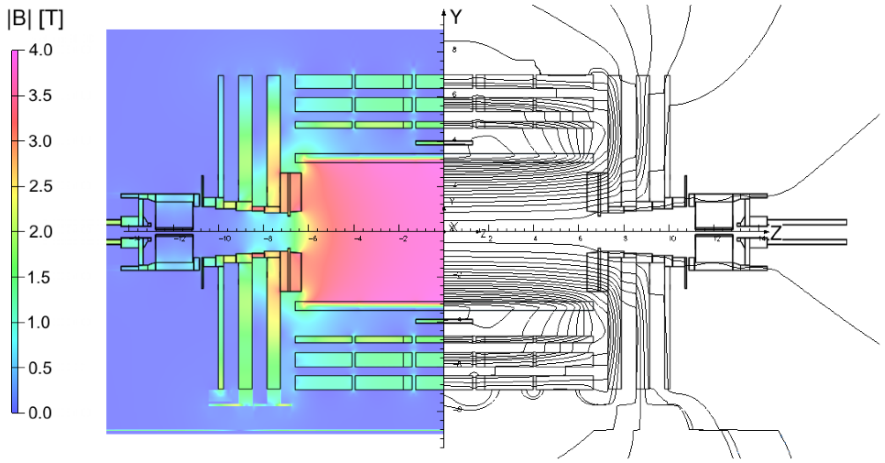


Figure 5.5: Simulated magnetic field strength and field lines of the CMS magnetic system Central solenoid and return yokes are visible.

the wear of the coil [8]. The increase up to nominal 4T will be decided when the aging of the magnet will be better understood.

The simulation of the magnetic field using finite-element numerical method is shown on Figure. 5.5. Inside the superconducting coil the field is very strong and almost homogeneous – the field is know to a precision better than 0.1%, which is crucial for the precise measuremetns of the charged particle tracks in the tracker.

The quality of the magnetic field predition was cross-checked using cosmic muon events [8] and by direct measurements of the field using probes both inside the solenoid [9] and in yoke [10].

5.2 CMS Trigger system

The design goal of the LHC is to have a bunch-crossing rate of 40MHz at instantaneous luminocities of about $\sim 10\,\mathrm{nb}^{-1}\mathrm{s}^{-1}$. This leads to the enormous number of 10^9 collisions per second each producing roughly 1 MB of data.

Therefore, recording all the retrieved infomation is, first of all, impossible with modern hardware – the storage capability of today systems is about $O(100)$ MB/s at 100 Hz rates, that is many orders of magnitude larger than the required ones. Secondly, most of that data contains uninteresting inelastic

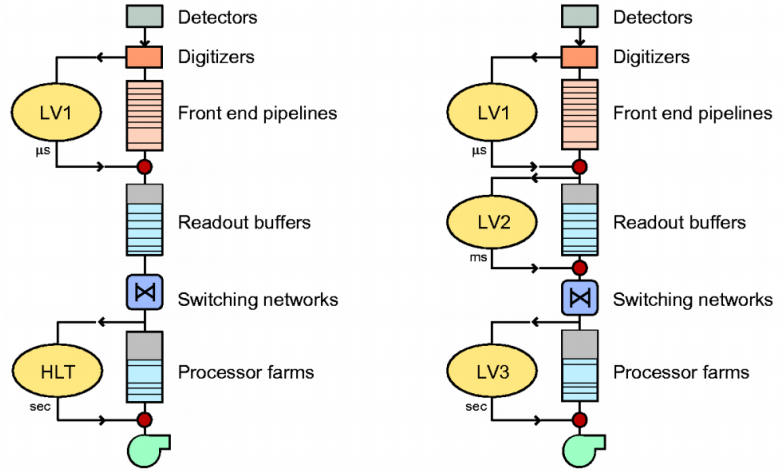


Figure 5.6: Two-level CMS trigger architecture (left), and conventional three-level trigger architecture (right).

pp collisons that hardly contain any information about the New Physics and can be easily discarded on-the-fly.

Trigger and Data Aquisition (TriDAS) systems of HEP particle physics experiments are designed to perform that kind of online data selection. The necessary six order of magnitude reduction in the event rate poses strict requrements on the TDAQ systems – the trigger's latency must be as minimal as possible, the event selection algorithms must be both effective at rejecting background processes and at keeping potential signal events even in such short times. Another important requirement lies in large amounts of data that should be transferred between detector front-end and various parts of the TDAQ subsystems.

The six order of magnitude reduction in rate from 40 MHz to 100 Hz is usually organized in several "trigger levels". The classic trigger system usualy contains three (or, sometimes even four) levels of data processing:

L1: first level triggert stays at the detector frontend and consists of the very low latency (about $\sim$ 1 ns) hardware electronics.

L2: level-2 is an itermediate level, made of programmable hardware that processes data at millisecond that is temporary stored in pipeline memory.

L3: final level-3 implemented as the computer farm running fully customizable software

As demonstrated of Figure 5.6, the CMS trigger design [11, 12] stands out in this repect – istead of investing in the control logic of the L2 layer, the decision was made to concentrate on increased bandwith and CPU processing power. As a result, CMS uses the two-level triggering system, combining conventional L2 and L3 in a single HLT stage:

- **The Level-1 (L1)** trigger uses programmable firmware electronics on Field Programmable Gate Array (FPGA) circuits. L1 employs simple selection algoririthms, optimised to reduce the event frequency from original 40 MHz to 100 kHz.
- **High Level Trigger (HLT)** is the second level implemented as a large CPU computing farm of about 13000 CPU cores, running the selection software written in C++ high-level programming language. It finalizes the selection providing another three orders of magintude reduction rate.

The L1 trigger uses only calorimetric and muon data – tracker information is unavailable at such a short times. The full tracker information is retrieved only when L1 acceptance is confirmed and the event is passed on to the HLT trigger.

At the HLT, timing is not as critical and all the detector data is present. HLT starts by reproducing L1 decision and then proceeeds by further filtering based on the analysis' needs – at the HLT level the trigger is able to perform numerous reconstructions and selections.

Bibliography

[1] Dan Green. "At the leading edge: The ATLAS and CMS LHC experiments" (2010). Ed. by Dan Green.

[2] CMS Collaboration. "CMS, tracker technical design report" (1998).

[3] CMS Collaboration. "CMS: The electromagnetic calorimeter. Technical design report" (1997).

[4] CMS Collaboration. "The CMS hadron calorimeter project : Technical Design Report" (1997).

[5] J. Freeman. "Innovations for the CMS HCAL". *Int.J.Mod.Phys.* A25 (2010), pp. 2421–2436.

[6] Christian Lippmann. "Particle identification". *Nucl.Instrum.Meth.* A666 (2012), pp. 148–172. arXiv:1101.3276 [hep-ex].

[7] CMS Collaboration. "The CMS magnet project : Technical Design Report" (1997).

[8] CMS Collaboration. "Precise Mapping of the Magnetic Field in the CMS Barrel Yoke using Cosmic Rays". *JINST* 5 (2010), T03021. arXiv:0910.5530 [physics.ins-det].

[9] V.I. Klyukhin et al. "Measurement of the CMS Magnetic Field". *IEEE Trans.Appl.Supercond.* 18 (2008), pp. 295–298. arXiv:1110.0306 [physics.ins-det].

[10] V.I. Klyukhin et al. "Measuring the Magnetic Flux Density in the CMS Steel Yoke" (2012). arXiv:1212.1657 [physics.ins-det].

[11] CMS Collaboration. "CMS. The TriDAS project. Technical design report, vol. 1: The trigger systems" (2000).

[12] CMS Collaboration. “CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger” (2002). Ed. by P. Sphicas.

Chapter 6

Physics objects reconstruction at CMS

The response of the detector is first interpreted and filtered by a complex trigger system and then stored in the data storage with intention to be further studied by various analyses. The detector response, therefore, should be interpreted in terms of Physics processes happening at the beam crossing and in the detector volume. Such interpretation must be done both during the data-taking (online), but it can also be redone with the recorded data (offline) to further improve the reconstruction quality by taking advantage of unbounded CPU time.

This very complex task is split into various subtasks: one starts with tracks and vertices reconstruction using tracker data. Stable SM particles are then identified using the whole detector information by means of the Particle Flow algorithm. Finally, aggregate objects like jets, τ-leptons and missing transverse energy are reconstructed. Development, evaluation and commissioning of the reconstruction algorithms (both online and offline) for all these subtasks are handled by specialized Physics Object Groups (POGs):

- **Tracking POG** maintains general tools for charged particle tracks reconstruction, determination of the beam-spot and the Primary Vetrtex.
- **B-tag and vertexing POG** mostly concerns b-tagging algorithms, but shares some responsibility on offline Primary Vertex determination.
- **Muon POG** works on tools for identification and reconstruction of

muons.

- **E/gamma POG** works on tools for identification and reconstruction of electrons and photons.
- **JetMET POG** is the largest POG, combining many activities related to jets – reconstruction algorithms, energy corrections and resolution, determination of missing transverse energy, e.t.c.
- **Tau POG** works on detection of τ-leptons both offline and online.

Together with the reconstruction algorithms, POGs also provide a number of identification criteria for their objects, namely some number of cuts on the object parameters (or some combined discriminator) are supplied for several "ID working points". Subsequent analyses might choose between the suggested selections, benefiting from reduced fake rates and better reconstruction and identification qualities at the expense of some reduction in reconstruction efficiency.

In this chapter a brief overview of the event reconstruction at CMS is presented with emphasis on the details that are most relevant for the analyses presented in the last part of the present work.

6.1 Track and vertex reconstruction

6.1.1 Track reconstruction

Track reconstruction is the first step in the whole event reconstruction procedure. The Kalman filter algorithm [1], first used in DELPHI experiment at LEP, is used as a basis for the track finding at CMS.

The reconstruction starts with "seeds": measurements in the pixel layers. The seeds are used as initial points for the iterative Combinatorial Kalman Filter processing [2]. This algorithm proceeds by iteratively constructing a tree of track candidates. For each consequent layer, hits that are compatible with previous layers predictions are added as branches to the data structure. Tree branches are cleaned by imposing some constraints on the number of sensor-crossings without hits and on the maximum χ^2 of the tracks' hits.

Description of the algorithm details and parameters at CMS are presented in [3]. The settings are chosen to have the best tracking efficiency with minimal fake rates. Another concern for the tracking algorithm is its CPU and

memory efficiency – the algorithms are very demanding even for offline reprocessing. In year 2011, with gradually increasing PileUp, numerous optimizations and logic improvements are performed [4], resulting in improved memory usage and factor of 7 speed-up of the processing time.

6.1.2 Vertex finding

Kalman filter technique can also be used for vertex finding. It turns out, though, that at high multiplicities the more robust algorithms like Adaptive Vertex Fit, outperform the simple Kalman filter approach [5] both with respect to speed and efficiency.

Reconstruction of the primary vertices is done using all of the event tracks. First the track compatibility with the beam spot (location of the LHC beam in the transverse plane) is checked by looking at the track quality and the distance of closest approach of the track to the beam line. The tracks are then clustered according to their z-coordinates at the points of the closest approach using deterministic annealing method [6].

To find the primary vertices the clusters are filtered by requiring a good fit quality and constraining the clusters to lie within $|z - z_0| \leq 24$ cm of the nominal detector centre z_0. Radial coordinate of the cluster must not be larger than $r \leq 2$ cm from the beamspot. The Primary Vertex with the largest $\sum p_T$ of the tracks is selected to be the event's hard interaction vertex. Remaining vertices on the beam axis are considered to come from PileUp interactions.

Detection of Secondary vertices (SV) – vertices at noticeable transverse distances from the beamline – is important for b-tagging and will be discussed in the next chapter.

6.2 PileUp reweighting

Beam intensities of the LHC was never encountered in HEP experiments before. Every collision generates many simultaneous interactions per bunch crossing, constituting the PileUp interactions. Figure 6.1 shows the observed distribution of interaction vertices per bunch crossing in 2012 $\sqrt{s} = 8$ TeV data.

All levels of the analyses at the CMS rely on numeorus Monte Carlo simulations of the detector. The results of such simulations strongly depend on

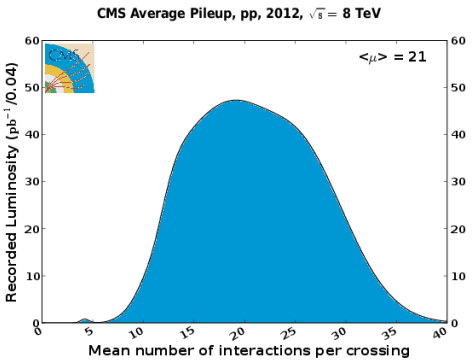


Figure 6.1: PileUp interactions per beam crossing in 2012 collisions.

the precise knowledge of the state of the detector, its readout and triggering electronics. Precise knowledge of the beam conditions are also required for the simulations to adequately reproduce the detector response.

As it will be discussed further, the PileUp interactions has an impact in a wide variety of measurements, related to the event reconstruction. The precise distribution on the Figure 6.1 is very hard to reproduce in Monte Carlo simulations. That is further complicated by changing beam conditions throughout the 2011-2012 years of data collection.

The standard method for overcoming the mismatch between simulated and observed distributions is to *reweight* the simulated samples as a function of the number of PileUp interactions. Reweighting is the common procedure when each simulated event gets a weight with which it contributes to the distributions of interest.

The reweighting tha accounts for data/MC mismatch in the number of PileUp interaction is done on event-by-event basis in such a way that the distribution in the number of reconstructed primary vertices in simulation matches this distribution in data. Figure 6.2 demonstrates the method for two analyses done in 2011 (see Chapter 9) and 2012 (see Chapter 10).

6.3 Particle Flow Algorithm

The next step of the event reconstruction at CMS is handled by Particle Flow (PF) algorithm [7–9]. It joins information from all subdetectors in the most op-

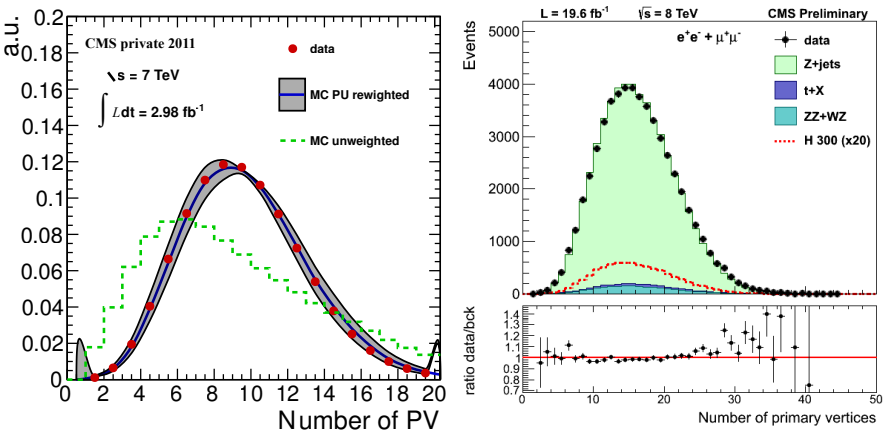


Figure 6.2: Examples of PileUp reweighting in 2011 (left) and 2012 (right) data.

timal way: maximizing reconstruction efficiencies, energy-momentum resolutions and minimizing the fake rates. PF algorithm returns a coherent list of reconstructed stable SM particles: electrons, muons, photons and both charged and neutral hadrons. The PF algorithm handles redundancies and simplifies corresponding isolation procedures – the list of returned particles should be as easy to use as the list of "true" particles in Monte Carlo simulations.

At the first step, PF collects all the identified "*elements*" from all the subdetectors – charged particles from tracker, clusters of energy deposits from ECAL and HCAL and muon tracks from the muon subsystem. Since a particle can leave several elements in those subdetectors, the Particle Flow attempts to associate the observed elements by "*linking*": track-HCAL, track-ECAL, ECAL-HCAL and track-muon segment links are considered.

The linking algorithm defines a distance for each pair of elements based on their compatibility. Tightly linked elements are then called a "*block*" – the set of elements representing a hypothetical particle. The linking takes into account physics properties of the detected particles and is optimized for individual Physics Object reconstruction.

The PF reconstructed particles are then used as ingredients for various energy-momentum correction and PileUp *mitigation* algorithms and as building blocks for reconstruction of combined objects: τ-leptons, jets and missing

transverse energy.

6.4 Jets and MET

LHC is essentially a hadronic machine: the protons it makes to collide are hadrons, made almost entirely of gluons with some amount of quarks. Gluons and quarks are the sole particles participating in the strong interaction, described by the Quantum Chromodynamics (QCD).

Being a well-established fundamental theory of strong interactions, QCD incorporates two notable phenomena: asymptotic freedom and confinement. Confinement forbids the separation of color-carrying quarks and gluons as isolated individual particles. Quarks and gluons only exist in form of color singlet hadrons – complex bound states with other strongly interacting particles .

The phenomenon of asymptotic freedom is an observation that the strong coupling constant becomes smaller with the renormalization flow at higher momentum transfers. This means that quarks and gluons inside hadrons interact more and more weakly as we zoom into smaller distances. Asymptotic freedom allows us to use perturbation theory and discuss the high-energy (hard) part of the collision in terms of individual quarks and gluons. The individual partons in the final state of the hard interaction are then continue into low-energy (soft) part of the collision, where formation of hadrons (called the quark *hadronization*) takes place.

Experimentaly, a hard parton with substantial p_T can be detected as a stream of hadrons, flying roughly in the same direction. Such a collection of hadrons is called a "*jet*". Jets allow us to reduce very complex events containing hundreds of particles into a simpler event picture, that contains crucial information about the underlying hard processes.

6.4.1 Jet Algorithms

There is no single strict definition of a "jet" – the jet reconstruction procedure depends on a number of details that could be selected depending on the preferences of a particular analysis. But it is customary to require that the jet algorithm satisfies the criterion of *infrared safety*: jet definition should be insensitive to collinear hadronizaiton effects, like gluon splitting or soft

gluon radiation. To describe the property more formally one considers a sequence $\mathcal{O}_n$ of observables each defined on a given number n of hadrons. The observables $\mathcal{O}_n$ are called infrared-safe if the following relations are satisfied:

$$\mathcal{O}_{n+1}(\dots, p, \dots) \xrightarrow{p\to 0} \mathcal{O}_n(\dots, \dots),$$

$$\mathcal{O}_{n+1}(\dots, p, q, \dots) \xrightarrow{p\| q} \mathcal{O}_n(\dots, p+q, \dots).$$

Infrared safety is important [10] because we are usually only able to calculate observables to a fixed-order in perturbation theory. Extension of such results to kinematic regions where validity of perturbation theory is questionable motivates the infrared-safety constraint.

Modern jet algorithms are clustering algorithms – they proceed by recursively combining pairs of "closest" fragments in a single "proto-jet" until all remaining distances do not exceed a certain predefined bound. The algorithm definition does, therefore, depend on a certain value R that restricts the size of maximal cluster in the space of the kinematic variables.

Another thing that should be specified for an algorithm is the very definition of the mentioned "closeness" of the tracks in the kinematic variable space. It is natural to require that such a measure should be (at least approximately) invariant under the Lorentz boosts along the beam axis. In the simplest case such a measure can be taken to be:

$$\Delta R_{ij} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2} \tag{6.1}$$

which exploits an important feature of rapidity – the difference between rapidities remains constant under boosts.

Apart from geometric separation between jet's constituents encoded in ΔR_{ij}, information about their energies should also be incorporated to influence the ordering of the clustering sequence, depending on the fragments energies. To achieve that, one modifies the measure (6.1) by introducing extra weights that depend on the fragment's kinematics.

A large class of the so-called k_t-based clustering algorithms can be described by the following expressions:

$$d_{ij} = \min((k_i^T)^{2p}, (k_j^T)^{2p}) \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = (k_i^T)^{2p} \tag{6.2}$$

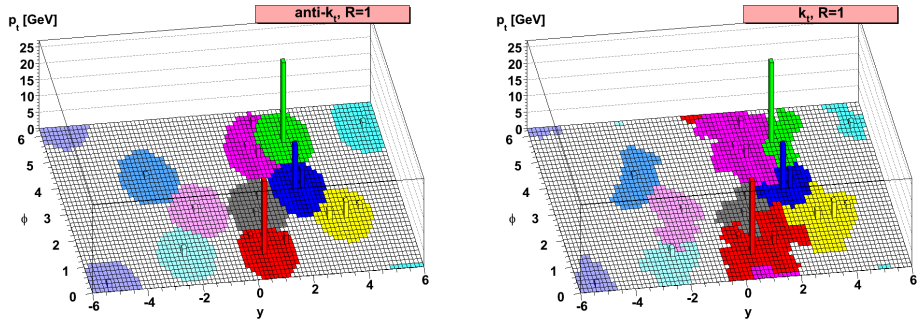


Figure 6.3: Comparison of anti-k_t (left) and k_t (right) jet clustering algorithms.

where the exponent p regulates the relative strength of energy (k^T) and geometry (ΔR_{ij}) contributions to the algorithm. The size of a jet cluster is bounded by introducing an extra distance d_{iB} between entry i and the beam B. If d_{iB} is the smallest distance among d_{ij}, then the fragment i is removed and called a final jet. The iterative procedure then continues until no entries remain in the list.

Based on the value (actually – only on the sign) of p, one retrieves several classes of the k_t-based algorithms:

- k_t **algorithm** is historically the first clustering algorithm, obtained from (6.2) by setting $p = 1$. The original motivation for a clustering algorithm was time-reversed recombination of the sequence of splittings in the parton shower – the softest and geometrically closest constituents should recombine first. This motivated the extra k^T weighting of the distance (6.2).

  The algorithm results in highly irregular jet edges because, by construction, softer constituents cluster earlier and form irregular patterns on jet boundaries. The jaggedness of the jet edges is considered to be a disadvantage as it complicates studies that generally rely on jet bounded geometry: detector acceptance, isolations, energy calibrations, e.t.c.

- **"Anti-k_t" algorithm** is the standard de-facto jet clustering algorithm used in analyses at LHC. It corresponds [11] to setting $p = -1$ (hence the "anti" prefix) in the (6.2) expression.

PF Jet ID	Loose	Medium	Tight
Neutral Hadron Fraction	< 0.99	< 0.95	< 0.90
Neutral EM Fraction	< 0.99	< 0.95	< 0.90
Number of Constituents	> 1	> 1	> 1
For $\|\eta\| > 2.4$ in addition			
Charged Hadron Fraction	> 0	> 0	> 0
Charged Multiplicity	> 0	> 0	> 0
Charged EM Fraction	< 0.99	< 0.99	< 0.99

Table 6.1: Jet ID criteria for PF jets

The "inverted" dependence on fragment k^T leads to the key feature of this algorithm – hard particles are combined in clusters before soft do. As a result, soft fragments are not affecting the jet's shape – only the hard ones do. The resulting jet boundaries become very circular, while the algorithm's infrared safety is still preserved.

- **Cambridge/Aachen (C/A) algorithm** is the special case $p = 0$ of the k_t-based algorithms. This algorithm is considered to be a "middle ground" between the two previous algorithms. On one hand it provides much more regular shapes of jet boundaries than k_t algorithm. Owning to the fact that C/A clustering depends only on angles between fragments, but not on theri energies.

  On the other hand C/A keeps one important advantage of the k_t algorithm– it preserves the clustering sequence inside the jets. This allows one to "revert" of the clustering sequence, providing a possibility to investigate jet's substructure and study various processes in boosted regimes.

6.4.2 Jet identification

Apart from the issues related to the representativeness of the jet, there are also problems associated with detector artifacts, electron fakes and jets originating from PuleUp interactions. The JetMET POG provides a recipe for dealing with such "fake" jets, by assigning a "jet quality" or jet identification to a particular jet.

High values of hadron and EM fractions of neutral particles are hints for unphysical jets from detector artifacts – Table 6.1 shows three working points

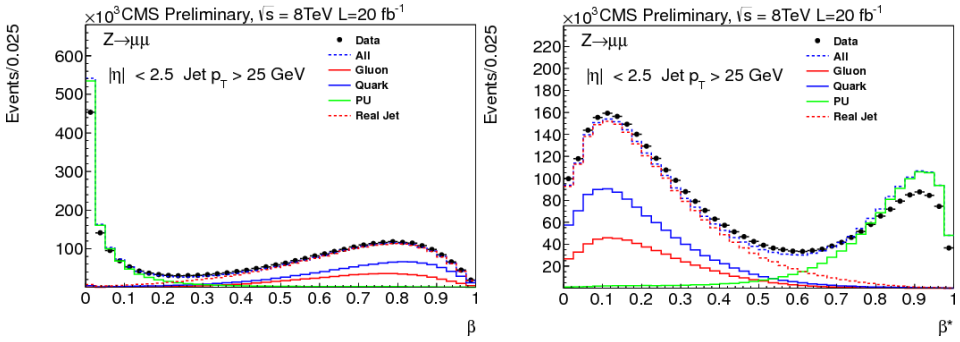


Figure 6.4: Simulation and data distributions of the two track-related PileUp-sensitive variables β and β^* for central jets with $p_T > 15$ GeV.

criteria recommended by JetMET POG for fakes removal. These criteria help with mitigation of instrumental artifacts and beam-related noise. But there is another concern to be dealt with: PileUp-originating jets, becoming an increasing problem as number of PileUp interactions grows. Jets from secondary vertices tend to "pollute" jets from the hard interaction vertex and also they might overlap with each other, combining into a single fake jet.

Identification of PileUp jets [12] relies on two major properties of such jets. First of all, the tracks of charged praticles tend to be incompatible with the Primary Vertex associated to the jet. Seconldy, PileUp jets are usually made of several smaller jets merged together, affecting their shape.

The two very useful track-based variables decribing a given jet are β and β^*, defined as p_T-weighted fraction of charged tracks that are coming from the jet's Primary Vertex (or from other PV in case of β^*):

$$\beta_{jet} = \frac{\sum\limits_{\substack{charged \in jet \\ from\ PV}} p_T}{\sum\limits_{charged \in jet} p_T}, \qquad \beta^*_{jet} = \frac{\sum\limits_{\substack{charged \in jet \\ other\ PV}} p_T}{\sum\limits_{charged \in jet} p_T}. \tag{6.3}$$

Figure 6.4 shows the distributions in the variables for data and simulation. The variables are very good at discrimination of PileUp jets.

The β and β^* variables are combined together with a number of other variables, related to jet's charged tracks (like z-distance of closest approach) and related to jet's shape (like p_T-wieghted variance of jet tracks' ΔR ditribution). All these varables are used to train a Boosted Decision Tree multivari-

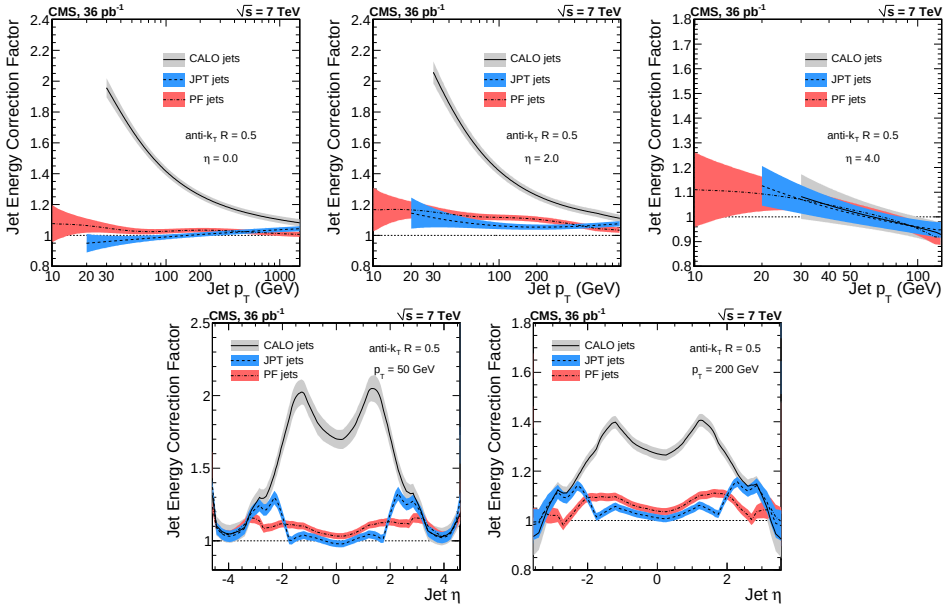


Figure 6.5: Central values and uncertainties of the combined Jet Energy Correction factors as a function of the jet's transverse momentum and pseudorapidity.

ate (the `PileUpMVA`) discriminator, efficient in the PileUp jet rejection in the whole pseudorapidity coverage $|\eta| < 5$ of the detector.

Moreover, for the tracker acceptance region $|\eta|$ the β and β^* variables are the most discriminating variables among others. Signal efficiencies of $\sim 99\%$ are reached for $85 \div 95\%$ PileUp jet rejection depeding on jet's p_T.

6.4.3 Jet Energy Scale Corrections

Being quite a complex objects, usually aggregating many soft particles, jets have an issue with their energy resolution. Even more concerning is that the measured jet's energy can be biased with respect to its true energy. Out-of-cone particles, noise in the detector, cracks and dead material are all contribution to the deterioration of the recorded jet's energy.

Jet Energy Corrections (JEC) are introduced [13, 14] in order to compensate for this bias. The correction is applied as a simple multiplicative factor

to the jet's four-momentum p_μ^{raw}:

$$p_\mu^{corr} = \mathcal{C} \cdot p_\mu^{raw} \tag{6.4}$$

The correction factor $\mathcal{C}$ is itself a combination of several multipliers, each depending on the corrected momentum from previous iteration:

$$\begin{aligned} p_\mu^{corr} &= \mathcal{C}_{abs}(p_T'') \cdot p_\mu'' \\ p_\mu'' &= \mathcal{C}_{MC}(p_T', \eta) \cdot \mathcal{C}_{rel}(\eta) \cdot p_\mu' \\ p_\mu' &= \mathcal{C}_{offset}\,(p_T^{raw}) \cdot p_\mu^{raw} \end{aligned} \tag{6.5}$$

Here $\mathcal{C}_{offset}$ is the offset correction that removes the extra energy that is not associated to the hard scattering part of the collision. Such additional contribution comes from the detector noise and from PileUp. This correction is evaluated using Jet Area, Average Offset or combined Hybrid Jet Area methods [13].

The $\mathcal{C}_{MC}$ factor corrects for non-uniformity in azimuthal angle and for non-linear response dependence on p_T. It is obtained from Monte Carlo simulated QCD events. Simulated events get reconstructed in the same way as it is done in actual data. Reconstructed jets are then matched to MC-true jets within $\Delta R < 0.25$ and the correction factor is extracted as an average fraction of p_Ts of the generated and reconstructed jets.

The Monte Carlo calibration $\mathcal{C}_{MC}$ accounts for the bulk of the reconstruction-MC mismatch. The last residual corrections $\mathcal{C}_{abs}$ and $\mathcal{C}_{rel}$ account for the differences between data and simulation, associated to imbalance in energy resolution as it depends on η (absolute correction) and p_T (relative corrections).

Figure. 6.5 shows the combined JEC factor (6.4) together with its estimated uncertainties for anti-k_t reconstructed jets with $\Delta R = 0.5$. Three types of jets are considered , depending on which information is supplied to the reconstruction algorithm: **calorimeter (CALO) jets** are jets reconstructed using only HCAL and associated ECAL deposits, **Jet-Plus-track (JPT)** are CALO jets with their energy-momentum improved by the charged tracks from tracker. Finally **Particle Flow (PF) jets** are jets reconstructed from the list of individual stable particles collected by the PF algorithm.

One can see that CALO jets require much larger corrections (especially at lower transverse momenta) because of the non-linearity of the calorimeters' response. Addition of the tracker information substantially reduces the necessary corrections in the tracker acceptance resion $|\eta| < 2.5$ – the transition

is especially sharp for JPT jets, while for PF jets it is much smoother due to some energy calibrations that was already applied for PF particles.

6.4.4 MET

Many analyses at the CMS involve the production of particles that penetrate whole detector's volume without leaving any trace in the active material. Such particles are therefore undetectable. Neutrinos are the SM particles that have this property, but it could also be other inert particles from some BSM extensions (usually these are candidates for the Dark Matter – see Chapter. 3.1.1).

Albeit undetectable directly, presence of such particles can be inferred by considering of the total transverse momentum of the event. The vectorial sum of the projection of the momenta of all particles on the transverse plane must be equal to (or very close to) zero, since it is true for the initial state of colliding proton pairs. Hence any considerable imbalance in the total transverse momentum would be a sign of some particles escaping the detector volume unobserved – such Missing Transverse Energy (MET) is determined as a sum of transverse energies of all observed particles in the event:

$$\not{E}_T = -\sum_i \vec{E}_{T_i}. \tag{6.6}$$

Missing Transverse Energy is another combined object, involving many particles of the event (actually, most of them) which makes it quite similar to jets. Jets and MET reconstruction are treated by the same POG, since many intricacies of $\not{E}_T$ determination are analogous to those of jet reconstruction.

As with jets, MET can be obtained using only calorimeter energies (Calo $\not{E}_T$), using tracks to improve the calorimeter only energy measurements (TC $\not{E}_T$), or by using particles, supplied by the Particle Flow algorithm (PF $\not{E}_T$). The PF technique shows very high performance [15, 16] and is used by majority of the analyses at the CMS.

Complementary to the Jet Energy Corrections, discussed in the previous section, the value of $\not{E}_T$ should also be corrected due to a variety of reasons.

- **Type-1 corrections** are applied by incorporating the jet energy corrections in the event:

$$\not{E}_T^{corr} = \not{E}_T - \sum_{jets} \left(\vec{p}_{T,j}^{corr} - \vec{p}_{T,jet} \right), \tag{6.7}$$

for PF $\not{E}_T$ the jets used in the sum in (6.7) must have $p_T > 10$ GeV.

- **Type-0 corrections** account for the charged-neutral particle imbalance in PileUp interactions – each PileUp vertex is expected to have well-balanced $\not{E}_T$, but this is not the case due to non-linearity of calorimeter response, showing a bias towards combined momentum of neutral particles.

$$\not{E}_T^{corr} = \not{E}_T - \sum_{vertices} f(|p_{T,v}^{ch}|) \frac{\vec{p}_{T,v}^{ch}}{|p_{T,v}^{ch}|}, \quad \vec{p}_{T,v}^{ch} = \sum_{charged} \vec{p}_T. \tag{6.8}$$

  Here $f(p_T)$ – is a function that parametrizes the correction that is obtained by simulation of single minimum-bias events.

- **The ϕ-asymmetry correction** is required to account for non-uniformity of the produced particles in the azimuthal angle. The cause of this asymmetry is not conclusively understood. It is present both in data and in simulation and can be attributed to relative misalignment of various detector subsystems or to a shift between the centre of the detector and the beamline.

As one can see, observed nonzero $\not{E}_T \neq 0$ might include resolution effects and uncertainties in the applied corrections. For a given event one might be interested in a quantity that reflects the consistency of the observed $\not{E}$

The MET significance is defined as the value of likelihood ratio test statistic (see equation (8.18) in the Section. 8.5) for null-hypothesis stating that $\not{E}_T^{true} = 0$ and alternative hypothesis claiming that $\not{E}_T^{true} = \not{E}_T^{obs}$:

$$\lambda_{MET} = 2 \ln \frac{P(\not{E}_T^{obs} | \not{E}_T^{true} = \not{E}_T^{obs})}{P(\not{E}_T^{obs} | \not{E}_T^{true} = 0)} \tag{6.9}$$

The MET significance λ_{MET} measures the certainty in the observed Missing Transverse Energy on the event-by-event basis. It can be especially useful for analyses that are interested in discrimination of events containing some MET without explicitly using its value.

Variable	2011 Cut	2012 Cut
Is `GlobalMuon`?	True	True
Is `PFMuon`?	–	True
$\chi^2/ndof$	< 10	< 10
N_{hits} pixel layers	≥ 1	≥ 1
N_{hits} tracker layers	≥ 9	≥ 6
N_{hits} muon chambers	≥ 1	≥ 1
$N_{matched}$ muon stations	≥ 2	≥ 2
Tracker $d_{x,y}$ distance to the PV	–	< 2 mm
Tracker d_z w.r. to the PV	–	< 5 mm

Table 6.2: Tight muon ID selection criteria for 2011 and 2012 analyses.

6.5 Lepton reconstruction at CMS

6.5.1 Muon reconstruction

Tracking of muons are performed at the CMS simultaneously by the inner silicon tracker and by gas-filled chamber detectors in the outer muon system. The muon tracks identified by the muon system are called *standalone-muon* tracks and the muon tracks in the tracker are called *tracker muons*. The so-called `GlobalMuon` is the track that is reconstructed using both the tracker and the muon system. If several tracker tracks can be matched to the standalone muon, then the muon with the lowest χ^2 is chosen to be a `GlobalMuon`.

The sources of the muon fakes could be: accidental matches, "hadronic punch-throughs", in-flight μ decays, tracks from cosmic muons and PileUp tracks. In order to reduce the fake rates and to guarantee good p_T measurement a number of Muon ID criteria was provided by the Muon POG.

Loose muons are generic muons reconstructed by PF algorithm; they could be either GlobalMuons or just a tracker muons (only Standalone Muons are very unlikely).

Soft muons are identified as tracker muons that are matched (inside-out) to at least one hit in muon system.

Tight muons are results of combined fit of tracker and muon system tracks. Good quality of the overall track fit is required.

Analyses presented in this work are relying on tight muon ID selection criteria. The details of these requirements used in 2011 and 2012 data are

	$\lvert\eta_{sc}\rvert < 1.479$			$1.479 < \lvert\eta_{sc}\rvert < 2.5$		
	Loose	Medium	Tight	Loose	Medium	Tight
$\lvert\Delta\eta_{trk}\rvert$	0.007	0.004	0.004	0.009	0.007	0.005
$\lvert\Delta\phi_{trk}\rvert$	0.15	0.06	0.03	0.10	0.03	0.02
$\sigma_{i\eta i\eta}$	0.01	0.01	0.01	0.03	0.03	0.03
H/E	0.12	0.12	0.12	0.10	0.10	0.10
d_0	0.02	0.02	0.02	0.02	0.02	0.02
d_Z	0.2	0.1	0.1	0.2	0.1	0.1
$\lvert 1/E - 1/p\rvert$	0.05	0.05	0.05	0.05	0.05	0.05
S^e_{Iso}/p^e_T	0.15	0.15	0.10	0.15	0.15	0.10
$N^{hits}_{missing}$	1	1	0	1	1	0
P^{vertex}_{fit}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}

Table 6.3: Cut-based electron ID criteria (for electrons with $p_T > 20$ GeV).

shown on the Table 6.2.

6.5.2 Electron reconstruction

Being very light, electrons quickly loose their energy in the strong CMS magnetic field due to Bremsstrahlung. Therefore electrons can be distinguished by relatively short tracks exhibiting the energy loss together with the corresponding ECAL cluster of Bremsstrahlung photons.

The so-called "ECAL-driven" electron reconstruction starts by identifying ECAL "super-cluster" – an energy deposit spread in an angular window $\Delta\eta \times \Delta\phi = 0.09 \times 0.6$. The cluster is then used as a starting point for dedicated electron track reconstruction – For more precise accounting for the energy loss, the Gaussian Sum Filter (GSF) fitting procedure[17] is used for electron tracks.

The procedure outlined above works well for high-pT electrons, but suffers from inefficient ECAL seeding for the particle transverse momenta lower than 10 GeV – the curvature radius of electron's trajectory gets smaller, spreading Bremsstrahlung photons over wider angles. The PF approach for electron reconstruction was designed with an intention to improve on this issue. The algorithm test all the tracks, compatible with electron hypothesis and associating independently reconstructed ECAL clusters. The associated supercluster in the ECAL is sought by taking all the tangents to the electron track at points of its intersection with tracker layers and extrapolating them to the calorimeter. As a result, the Particle Flow algorithm is able to reconstruct

both high-p_T and low-p_T electrons almost independently on their isolation.

For the electron ID, the e/γ POG provides several identification techniques:

- **Simple cut-based electron ID** is a robust and transparent procedure for analyses that do not need the ultimate electron efficiency.
- **MVA-based electron ID** combines all relevant data from the HCAL, ECAL and tracker, resulting in up to 30% efficiency improvement compared to simple cut-based identification.
- **HEEP electron ID** targets high-energy electron pairs (HEEP) identification for vaious exotica searches and usese only measurements from calorimeter clusters.

The cuts for the simple cut-based ID are shown on Table 6.3. One can see that the cuts are different for the barrel $|\eta| < 1.479$ and endcaps $1.479 < |\eta| < 2.5$ pseudorapidity regions.

The discriminating variables $\Delta\eta_{trk}$, $\Delta\phi_{trk}$ are measuring the quality of angular matching between supercluster and the track. The $\sigma_{i\eta i\eta}$ describes the shape of the supercluster along the asimutal direction. The H/E is the ratio between energy deposits in HCAL tower behind the supercluster and the energy of the supercluster itself, while $|1/E - 1/p|$ tells the difference between ECAL-measured energy and tracker-measured momentum. Finally, S^e_{Iso}/p^e_T is the relative electron isolation, that we are going to describe below.

6.5.3 Lepton isolations and efficiencies

In the framework of the Particle Flow algorithm, (absolute) lepton isolation is defined as a sum of transverse momenta of hadrons (both charged and neutral) and photons inside a certain ΔR cone around the lepton momentum. The isolation helps in discriminating leptons that arise from QCD processes from those appearing from decays of heavy particles (like Z or W).

Isolation is the quantity obtained from the summation over soft PF particles in a certain angular region around the lepton. As we have seen in cases of jets and MET, such quantities are sensitive to contamination from PileUp-originating particles. Again, some correction should be applied on the observed isolation to mitigate that effect – the value is derived [18, 19] from the average PileUp energy density ρ evaluated on event-by-event basis and the

effective cone area A_{eff} in angular space. The final expression for PileUp-correctted absolute isolation reads:

$$S^{\ell}_{Iso} = \sum_{charged} p_T + \max\left(0, \sum_{neutral} E_T + \sum_{\gamma} E_T - \rho \cdot A_{eff}\right). \tag{6.10}$$

Electron and muon isolation cuts are then usually imposed not on absolute isolation (6.10), but on the realtive isolation $S^{\ell}_{Iso}/p^{\ell}_T$. The POG-recommended values for muon isolation are:

$$\Delta R = 0.4, \qquad S^{\mu}_{Iso}/p^{\mu}_T < 0.12, \tag{6.11}$$

for the isolation cone size and tight relative isolation cut. Similar values for electrons are:

$$\Delta R = 0.3, \qquad S^{e}_{Iso}/p^{e}_T < 0.15. \tag{6.12}$$

The efficiencies of the lepton reconstruction (including identification, isolation and trigger efficiency) can be obtained from data using the so-called *tag-and-probe* method [20]. The procedure starts with selection of $Z \to \ell^+\ell^-$ events. Events are required to have exactly two lepton candidates with invariant mass, fitting the $m_{\ell\ell} \in [60, 120]$ GeV window. One of the leptons – the so-called "*tag*" – is required to pass tight selection and isolation requirements. Second lepton from the pair – the "*probe*" – is selected using whatever criteria are being studied.

The signal yields for accepted and rejected probe leptons are then extracted by fitting with the invariant mass distribution for the di-lepton system and the selection efficiencies are calculated as the relative signal yields of accepted and rejected probe leptons. This way the efficiency is measured in different bins of the $\eta \otimes p_T$ space.

Bibliography

[1] R. Fruhwirth. "Application of Kalman filtering to track and vertex fitting". *Nucl.Instrum.Meth.* A262 (1987), pp. 444–450.

[2] R. Mankel. "A Concurrent track evolution algorithm for pattern recognition in the HERA-B main tracking system". *Nucl.Instrum.Meth.* A395 (1997), pp. 169–184.

[3] W. Adam et al. "Track reconstruction in the CMS tracker" (2005).

[4] D. Giordano and G. Sguazzoni. "CMS reconstruction improvements for the tracking in large pile-up events". *J.Phys.Conf.Ser.* 396 (2012), p. 022044.

[5] Wolfram Erdmann. "Vertex reconstruction at the CMS experiment". *J.Phys.Conf.Ser.* 110 (2008), p. 092009.

[6] E. Chabanat and N. Estre. "Deterministic annealing for vertex finding at CMS" (2005), pp. 287–290.

[7] CMS Collaboration. "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET" (2009).

[8] CMS Collaboration. "Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector" (2010).

[9] "Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV" (2010).

[10] Gavin P. Salam. "Elements of QCD for hadron colliders" (2010). arXiv:1011.5131 [hep-ph].

[11] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The Anti-k(t) jet clustering algorithm". *JHEP* 0804 (2008), p. 063. arXiv:0802.1189 [hep-ph].

[12] CMS Collaboration. "Pileup Jet Identification" (2013).

[13] CMS Collaboration. "Plans for Jet Energy Corrections at CMS" (2008).

[14] CMS Collaboration. "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS". *JINST* 6 (2011), P11002. arXiv:1107.4277 [physics.ins-det].

[15] CMS Collaboration. "Missing transverse energy performance of the CMS detector". *Journal of Instrumentation* 6 (Sept. 2011), p. 9001. arXiv:1106.5048 [physics.ins-det].

[16] "MET performance in 8 TeV data". CMS-PAS-JME-12-002 (2013).

[17] Wolfgang Adam et al. "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC" (2005).

[18] Matteo Cacciari and Gavin P. Salam. "Pileup subtraction using jet areas". *Phys.Lett.* B659 (2008), pp. 119–126. arXiv:0707.1378 [hep-ph].

[19] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The Catchment Area of Jets". *JHEP* 0804 (2008), p. 005. arXiv:0802.1188 [hep-ph].

[20] CMS Collaboration. "Measurements of Inclusive W and Z Cross Sections in pp Collisions at $\sqrt{s} = 7$ TeV". *JHEP* 1101 (2011), p. 080. arXiv:1012.2466 [hep-ex].

Chapter 7

B-tagging at CMS

In many analyses the information about the type of parton generating hadronic jets plays a major role. Such information is supplied by the so-called jet tagging algorithms. The algorithms produce a discriminator value for a given jet. The value of the discriminator can roughly be interpreted as a likelihood of the jet to have a particular parton as its origin. One can then cut on the value of such a discriminator, choosing the best purity-to-efficiency ratio for his needs.

The most attention is paid to the "b-tagging" – tagging of the jets originating from the b-quark. The reason for such acclaim is based primarily on the efficiency of such tagging, which in turn rests on specific features of the b-quark, like its heavy mass and large lifetime of the hadrons containing it (b-flavored hadrons).

7.1 Properties of the b-quark

Mass of the b-quark is $m_b = 4.18\text{GeV}$. It is the second (after the t-quark $m_t = 160\text{GeV}$) heaviest fundamental fermion of the Standard Model. The t-quark has a really short lifetime – so short that it doesn't form any hadrons at all. The b-quark, on the other hadn does form hadrons combining with even lighter partner quarks: u, d or s.

Having a comparatively large mass, b-quarks are typically produced with noticeable transverse momentum during the hard interaction phase. During the hard phase pair-produced b-quarks decay independently, thus originat-

ing separate hadronization. Throughout the soft hadronization phase the hadron that incorporates a b-quark is the most energetic among its "neighbors". The typical multiplicity of the charged products in b-hadron decays is about ~ 5.

The most prominent feature of the b-flavored hadrons is that their ground states decay only via weak interactions. The b-quark decay is dominated by the $b \rightarrow cW^*$ process with virtual W^* creating either a pair of quarks (hadronic decay) or a lepton-neutrino pair (semileptonic decay).

Since b-hadrons decay only weakly, and due to the smallness of the corresponding CKM parameter $|V_{cb}|$, their lifetime is many orders of magnitude longer than that of many "ordinary" hadrons. With $c\tau \sim 450\mu m$ one can have the b-hadron decaying considerably far (up to several millimeters) away from the primary interaction vertex.

Combining that with already mentioned properties of large mass and transverse momentum, one expects to have a secondary vertex substantially far away from the collision axis, with large invariant mass at that vertex.

Finally one should note that b-hadrons have high semileptonic branching ratio, which means that quite often there are leptons inside the jet, coming from the mentioned secondary vertex.

7.2 B-tagging algorithms

B-tagging algorithms summarize some of the jet data relevant for b-quark identification in a single discriminator value. By convention larger values of the discriminator mean larger likelihood of the jet to be a b-jet. Apart from that, there are no extra conventions on the behavior of the discriminators: they are not necessary fixed to a particular range and are not necessary smooth – distributions in them could have various discontinuities, spikes and other kinds of irregularities.

The most popular algorithms at CMS (recommended for the analyzes by the b-tagging working group) are Track Counting, Jet Probability and Combined Secondary Vertex algorithms. Figure 7.2 shows the discriminator efficiencies for various modification of these algorithms, discussed further in the chapter. It is clear from the plots that b-tagging can achieve a considerable discrimination efficiency especially for light quarks and gluons, with the largest source of the fakes being the c-quark originating jets.

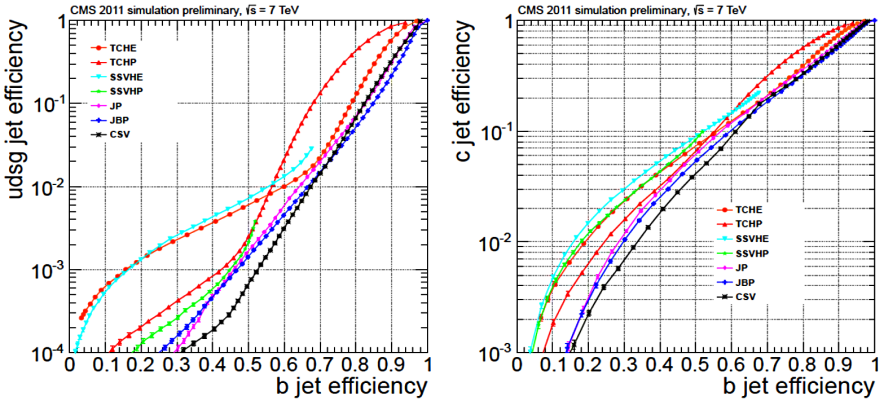


Figure 7.1: B-tag discriminator perfomances for various jet types evaluated from Monte Carlo : b v.s. $light$ (left) and b v.s. c (right) efficiencies are shown.

Every b-tagging algorithm gets as an input a jet, a collection of tracks associated with it and the position of the jet primary vertex (PV). B-tagging algorithms rely heavily on the relative spatial positions and orientations of these objects. Also, algorithms that involve reconstruction of the secondary vertex are faced with huge combinatorial complexity of the vertex finding procedure. This is further complicated by vertexes from PileUp interactions.

That is why additional track selection requirements are imposed on each of the jet tracks: to ensure a certain quality of the jet-associated tracks and to reduce the number of tracks for vertex reconstruction, they are:

- track should lie close to the axis of a jet under consideration
  - $\Delta R < 0.3$ if SV-reconstruction is needed
  - $\Delta R < 0.5$ otherwise
- $p_T > 1$ GeV and $\chi^2/ndof < 5$,
- there must be at least eight tracker hits, among them at least two hits in the pixel tracker,
- extra geometric requirements to remove fakes from other long-lived particles

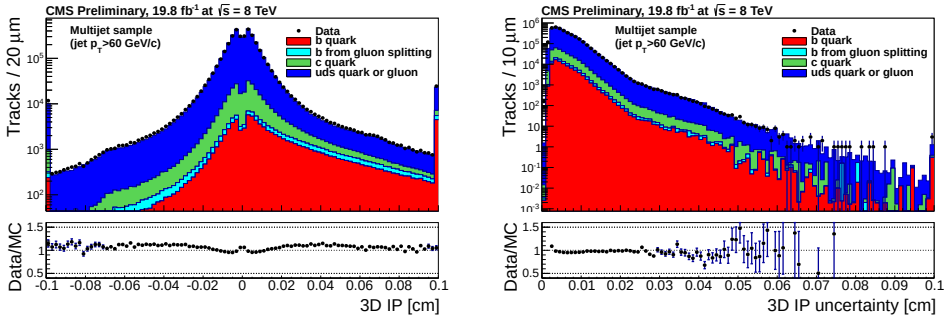


Figure 7.2: 3D impact parameter (left) and the 3D-IP uncertainty for an inclusive multijet ("QCD") sample. Various kinds of jets are presented as different colors of stacked histograms.

7.2.1 Track counting b-tagging

Most of the b-tagging algorithms exploit the long-lifetime property of b-flavored hadrons. To do so, it is necessary to obtain certain information about the point, where the long-lived b-hadron did decay. Simple observable, associated to the individual tracks and containing this kind of information is the impact parameter (IP).

Impact parameter value is defined as the minimal distance at which the track comes to the primary vertex. The value of IP can be positive or negative based on the relative direction of the jet axis and the IP-segment. One can also consider either full 3D reconstructed impact parameter or consider only its 2D projection on the plane transverse to the collision axis.

It is interesting to note that the expected IP value is independent of the decaying particle kinetic energy: while the lifetime and, therefore, average travel distance of the decaying particle increases by the relativistic γ factor – it gets compensated by the fact that the typical decay angle of the products of the boosted system are concentrated around the flight axis, with the typical angle scaling as $\sim 1/\gamma$.

Fig.7.2 shows the distributions in the IP and the corresponding measurement uncertainty [1, 2]. One can see that the impact parameter does have a certain uncertainty associated with it, and that uncertainty can be comparable to the measured IP value. To that end, it is more useful to utilize the impact parameter significance ($S_{IP} = IP/\sigma_{IP}$) instead of the "plain" IP.

Track counting algorithm provides one with discriminators by first sorting the jets' tracks by their IP-significance values and then returning the IP-significance for the N'th jet in that sorted list. Conventional values for N are: $N = 2$ for the so-called Track Counting High Efficiency discriminator and $N = 3$ for the Track Counting High Purity discriminator.

7.2.2 Jet probability b-tagging

Track counting algorithms are based only on IP-significance value and only return this value for just a single track in a given jet. Although very straightforward, there is clearly some room for improvement here. One can get more complex but more powerful discriminator by considering extra information about individual tracks and by combining the information from all the tracks in a jet.

Jet probability algorithm does this by evaluating probabilities P_i for each of the jet's tracks $i = 0, 1 \cdots N-1$ to originate from the primary vertex. The distributions for that evaluation are precomputed from a sample of PV-originated tracks. One can easily obtain such a sample even from data by considering tracks with negative IP: if we also assume that the distribution for the tracks from PV is the same for $S_{IP} > 0$ as for $S_{IP} < 0$, then it is possible to pre-compute the distributions in $|S_{IP}|$ from data, and :w utilize them in the analysis.

Describing the procedure in detail: one first subdivides all the tracks in the sample into several classes according to various track variables: p, $|\eta|$, number of hits in the pixel detector and the track's $\chi^2/ndof$. Then for each class one constructs histograms in the track's $|S_{IP}|$. These histograms are then interpreted as probability density functions of PV-originating track to have that value of $|S_{IP}|$.

During the analysis, the track probability P_i is defined [3] as a statistical significance of the observed $|S_{IP}|$:

$$P_i = \int_{|S_{IP}|}^{\infty} PDF(s_{IP}) ds_{IP}$$

Next step is to combine the probabilities of the individual tracks into a single Jet Probability (JP) discriminator:

$$D_{JP} = -\frac{1}{4} \log \left(\Pi \sum_i \frac{(-\ln \Pi)^i}{i!} \right), \quad \Pi = \prod_i max(P_i, 0.005). \tag{7.1}$$

The expression (7.1) represents a standard combination of confidence levels for all the tracks to be consistent with the primary vertex. Notice that probabilities P_i are bounded from below by the small probaility of 0.5%. This is done because the tracks with very small values of P_i tend to "overshadow" tracks with larger P_i reducing the disrimination power of the tagger.

Jet B Probability [4] is a slight modification of the previous discriminator, that exploits our knowledge of charged multiplicities of the b-quark decays. The discriminator (7.1) is modified as:

$$D_{JBP} = -\frac{1}{4}\log\left(\Pi\sum_{i<4}\frac{(-\ln\Pi)^i}{i!}\right), \quad \Pi = \prod_i max(P_i, 0.005). \tag{7.2}$$

The cut $i < 4$ on the number of leading tracks is motivated by ~ 5 charged multiplicity of the b-hadron decays and by the $\sim 80\%$ Impact Parameter reconstruction efficiency.

7.2.3 Simple Secondary Vertex b-tagging

Jet Probability and Track Counting algorithms are based on the impact parameter values of individual tracks. The discrimination power can be further improved by attempting to reconstruct the secondary vertex, produced by the decay of the b-flavored hadron.

Reconstruction of the secondary vertexes is performed by the adaptive vertex fit algorithm. Given a set of tracks, the algorithm returns a candidate for a secondary vertex together with weights $w_{tr} \in [0, 1]$ which measure the consistency of each track with that vertex candidate. One then proceeds iteratively, by removing tracks with $w_{tr} > 0.5$ from the tracks collection and calling the adaptive vertex fit algorithm again. These steps are repeated until the vertex reconstruction fails.

All the newly reconstructed secondary vertex candidates are then required to pass extra criteria. The SV candidate is removed if:

- more than 65% of tracks are shared with the PV,
- flight direction is too far ($\Delta R > 0.5$) from the jet's axis,
- transverse distance from PV is more than 2.5 cm or less than 0.1 mm
- invariant mass is compatible with K^0.

The Simple Secondary Vertex discriminator is then based on the significance of the 3D distance d_{fl} between the reconstructed SV and the primary vertex.

$$D_{SSV} = \log\left(1 + \frac{d_{fl}}{\sigma_{d_{fl}}}\right)$$

7.2.4 Combined Secondary Vertex b-tagging

The Simple Secondary Vertex algorithm relies upon the flight distance and number of tracks at the vertex. However there are more properties to consider: mass at the vertex, energies, p_Ts and ηs of the vertex's tracks, e.t.c.

CSV discriminator is targeted to combine all the possible information both from impact parameter and from secondary vertex in a single Multivariate Analysis (MVA) based discriminator. If the SV is not available the algorithm just combines two or more S^{2D}_{IP} tracks in into a so-called "pseudo vertex" . From this object it is then possible to obtain some of the vertex's variables without doing the actual vertex reconstruction. If even "pseudo vertex" reconstruction is impossible, then CSV falls back to the impact-parameter-variables-only mode.

Using all the attainable variables, two likelihood-ratio discriminators are pre-produced: D_c, that discriminates b-jets against c-jets, and D_{uds}, that discriminates against light-quark jets. Finally, CSV returns a single discriminator, combined as follows:

$$D_{CSV} = \frac{1}{4}(3D_c + D_{uds})$$

7.3 Online b-tagging in the HLT trigger

Requiring b-tagging at trigger level allows one to tackle with large rates at low p_T thresholds. Thus increasing sensitivity to various processes involving b-quarks and simultaneously reducing many backgrounds outright at the stage of online data acquisition.

Any algorithm used for the online data processing faces strong CPU constraints. For every particular process the balance between computational robustness and selection efficiency should be found. For b-tagging algorithms the CPU restraint is further tightened by the need of good identification of the primary vertex. Standard HLT implementations provide one with only a 1D

estimation of the vertex's position, while with increasing number of PileUp interactions, the algorithmic complexity of track end vertex reconstruction grows non-linearly.

The implementations that are the most relevant for the present work is the class of "`BTagIP`" implementations used in 2011 data collection. All of them are based on the online version of the Track Counting algorithm. The algorithm is applied on jets both at `L2.5`, where only the pixel tracks are available, and at `L3`, when the complete tracking is performed also by means of the tracker hits. A three-dimensional vertex is reconstructed in order to improve the performance of the algorithm and to ensure rate stability in case of poor beam-spot online definition. Several consequent modifications of this path were used by us:

BTagIP Performs TCHE b-tagging on the four leading central jets in the event. The values of S_{IP} are calculated in the 2D plane, transverse to the beam axis. The event passes the trigger if $S_{IP}^{2^{nd}track} > 3.5$.

BTagIP3D In that version, 3D primary vertex reconstruction is performed. The threshold on the significance of the second track is unchanged.

DiBTagIP3D1stTrack That path requires two jets to be b-tagged. At the same time the S_{IP} cut is now applied to their respective first tracks.

In 2012, the new "fast primary vertex" algoritm was implemented at HLT. The algorithm reconstructs the PV in two steps – first it reconstructs a "coarse" primary vertex is obtained by matching online calorimeter jets to pixel clusters and then projecting the clusters onto the z-axis. Then the precise PV position is obtained, considering only the tracks from the "coarse" vertex.

The algoritms improves the speed of vertex reconstruction by a factor of five. Such a large CPU time reduction allows one to use CSV class of algorithm for b-tagging at the HLT level.

Bibliography

[1] CMS Collaboration. "Performance of the b-jet identification in CMS" (2011).

[2] CMS Collaboration. "Performance of b tagging at sqrt(s) = 8 TeV in multijet, ttbar and boosted topology events" (2013).

[3] G. Borisov and C. Mariotti. "Fine tuning of track impact parameter resolution of the DELPHI detector". *Nucl.Instrum.Meth.* A372 (1996), pp. 181–187.

[4] "Algorithms for b Jet identification in CMS" (2009).

Chapter 8

Statistical analysis

The field of statistics is devoted to studying and interpreting data. Statistics is used in particle physics experiments to refine and present the results of the analysis. Probability theory serves as a formal mathematical foundation of statistics. In this chapter we present an overview of statistical methods used in the analysis described in this work, starting with a brief recap of the probability theory.

8.1 Probability theory

The bedrock for probability theory is the notion of *random variable*. Random variables do not have a fixed value and are subject to random variation in a certain set Ω of some possible values (like the set of possible outcomes of some experiment). Each random variable has an associated measure on the set of its outcomes – for every subset $\mathcal{F}$ of the sample space, there is a defined probability $P(\mathcal{F}) \leq 1$, which should satisfy certain axioms [1] to be a proper probability measure on that set.

If the variable is real-valued and one-dimensional, then the probability measure can be described by means of a distribution function or, more generally, a family of distribution functions, parametrized by (one or several) parameter θ:

$$F(x;\theta) = P(X \leq x|\theta). \tag{8.1}$$

Here we adopt standard notational conventions for probability and statistics: to distinguish random variables from "ordinary" numbers we will write

them in upper case like X. Particular realizations of a random variable will be written as corresponding lower-case letters. Greek letters like θ will be used for parameters of various distribution functions.

Equation (8.1) defines the so-called *cumulative distribution function* (*c.d.f*) for a random variable, which can be either continuous or discrete. If X is a continuous random variable, then it is possible to express the *c.d.f* as an integral of a *probability density function* (*p.d.f*):

$$F(x;\theta) = \int_{-\infty}^{x} f(x';\theta)\, dx' \,. \tag{8.2}$$

The equatoin reads "probability of X to be less than or equal to x, given some values of the parameters θ". If X is a discrete random variable, then its individual outcomes x_i have nonzero associated probabilities $P(X = x_i)$. This defines $p(x_i;\theta)$, a *probability measure function* (*p.m.f*) for X:

$$F(x;\theta) = \sum_{x_i \le x} p(x_i;\theta)\,, \quad p(x_i;\theta) = P(X = x_i|\theta). \tag{8.3}$$

Likelihood function is any parameter-dependent probability $P(\mathcal{F}|\theta)$ seen as a function of its parameters, given a particular outcome x of the corresponding random variable. Typically likelihood functions are constructed from *p.d.f*s (or *p.m.f*s) like:

$$L(\theta) = f(x;\theta). \tag{8.4}$$

It is obligatory to comment here, that $L(\theta)$ is not a *p.d.f* for θ, first of all because $L(\theta)$ is not normalized. More subtle question is: is it possible to consider θ as a random variable in the first place? This leads us to the examination of conceptual differences between so-called *frequentist and Bayesian* points of view on probability and statistics.

From *the frequentist point of view* statements are either true or false. The probability of a certain statement is the frequency with which true ones are present in a set of statements of similar nature. Following this line of thought we have to conclude that there must be a certain unknown but fixed "true value" of the measured parameter θ. So it cannot be considered as a random variable and *p.d.f* for θ simply cannot be defined.

Bayesian approach, on the other hand, considers probabilities as measure of our degree of belief in various statements. These probabilities are progressively updated by means of Bayes's formula. To demonstrate the bayesian

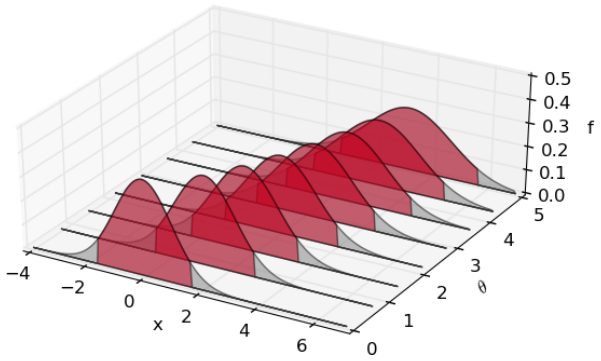


Figure 8.1: Construction of Neyman confidence belt

approach, consider the following example: let us define the statement $B =$ *"given jet is a b-jet"*, and the statement $T =$ *"b-tagging test is positive"*.

In that case $P(B)$ will be our "prior" – it is measuring our degree of beleif in the statement B. The probability $P(T|B)$ is the "posterior" probability – it measures our degree of beleif in B after we've observed T. The Bayes Theorem reads:

$$P(B|T) = \frac{P(T|B)}{P(T)} \cdot P(B), \tag{8.5}$$

where the factor $\frac{P(T|B)}{P(T)}$ is called a "Bayesian factor", representing the strength of eveidence that the observation T gives us in favor of the hypothesis B.

Systematic application of the "Bayesian update" procedure (8.5) is the basis of Bayesian point of view on probabilities: any statement can have a probability associated to it as long as it can be changed by means of the Bayes Theorem. From that point of view parameters like θ can be considered as random variables and do have probability distributions associated to them.

8.2 Neyman confidence belts

Given a number $1 - \alpha$, the Neyman's *confidence interval* for the random variable X with the probability distribution fucntion $\mathrm{pdf}(x)$ is any interval $[x_1, x_2]$, that satisfies:

$$P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} \mathrm{pdf}(x)\, dx = 1 - \alpha. \tag{8.6}$$

The number $1 - \alpha$ is then called a *confidence level* (CL) for such intervals. If X is a discrete random variable then it is sometimes impossible to find x_1, x_2 for equality (8.6) to be exactly satisfied. For discrete random variables the definition (8.6) is conventionally relaxed to:

$$P(X \in [x_1, x_2]) = \sum_{x_i \in [x_1, x_2]} \text{pmf}(x_i) \geq 1 - \alpha, \tag{8.7}$$

where $\text{pmf}(x)$ is the probability measure function for the discrete random variable X.

Both (8.6) and (8.7) are ambiguous – there is usually a continuum of x_1, x_2 pairs, satisfying these equations. Some extra criteria for x_1 and x_2 must be chosen. For example, one might choose $x_1 = -\infty$, thus getting an interval bounded above by x_2 – the so-called "upper confidence limit":

$$P(X \in [-\infty, x_2]) = P(X < x_2) = 1 - \alpha. \tag{8.8}$$

Another natural definition is to require equal probabilities of X being on the each side of the interval:

$$P(X < x_1) = P(X > x_2) = \frac{1 - \alpha}{2}, \tag{8.9}$$

creating the so-called "central confidence intervals".

Suppose that we have an experiment that measures some quantity described by a random variable X distributed according to one of the probability distribution functions from a family $f(X; \theta)$. The true value of θ is unknown and the goal is to provide an interval estimate of the value of $\theta \in [\theta_1(x), \theta_2(x)]$ for a given outcome x.

Figure 8.1 describes, following [2], a construction of the so-called *confidence belt*. For every value of θ we can substitute the corresponding *p.d.f* into (8.6) and obtain the confidence interval $[x_1(\theta), x_2(\theta)]$. The resulting region of the $x \times \theta$ plane is bounded by the functions $x_1(\theta)$ and $x_2(\theta)$.

Observe now that these boundaries can also be considered as functions $\theta_1(x)$ and $\theta_2(x)$ and that the following equality holds:

$$P(x_1(\theta) < X < x_2(\theta)) = P(\theta_1(X) < \theta < \theta_2(X)) = 1 - \alpha \tag{8.10}$$

Let us stress that θ in the expression (8.10) is not a random variable – this is crucial for the frequentist interpretation of confidence belts. If a number

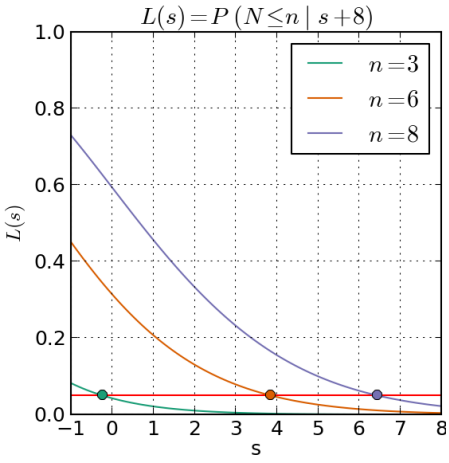


Figure 8.2: Poisson likelihood functions for a counting experiment with $b = 8$. Upper bounds at 95% CL for various outcomes are shown.

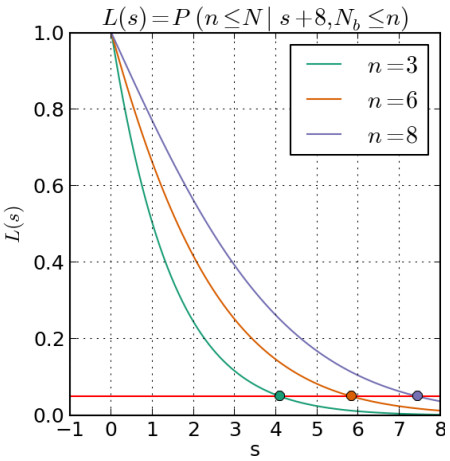


Figure 8.3: The "renormalized" likelihood function with the corresponding upper limits on s.

of experiments measuring X are performed and for each experiment the constructed confidence interval $[\theta_1, \theta_2]$ is published, then α is the fraction of the reported confidence intervals that do not contain the true value of θ.

8.3 Small signals problem

Let us now consider a simple event-counting experiment: suppose that there are signal and backround processes contributing to the observed number of events N in some experiment. The signal and background events are occuring independently with average rates s for the signal and b for the background.

Suppose also that there is no way we can distinguish signal and background events – the only observable we have access to is the number of events N. In that case N can be considered as a discrete random variable, characterized by the Poisson probability distribution with average rate equal to the sum $s + b$ of the signal and background rates:

$$P(N = n|s, b) = Poisson(n; s + b) = \frac{(s+b)^n}{n!} \cdot e^{-(s+b)}\,. \tag{8.11}$$

Assuming that the expected rate of the background events b is known, we would like to set an upper limit on s, given an observed number of events n. Naively one would use the technique for construction of confidence intervals described in the previous section. Using the prescription (8.7) for construction of confidence intervals for a discrete random variable, and following the procedure (8.10) one obtains:

$$L(s) = P(N \leq n|s, b) \leq \alpha\,, \tag{8.12}$$

it says which values os s variable are excluded with confidence $1 - \alpha$ is we ahve observed n events.

Figure 8.2 shows several likelihood functions for $b = 8$ and different values of n. The $L(s) < 0.95$ bound is drawn as a red line and the boundaries of the one-sided confidence intervals are marked as circles. For example, if the expected background rate is $b = 8$ and we've observed $N = 6$ events, then the values of s above $s > 3.8$ are excluded at 95% confidence level.

One can also spot the problem that Fig.8.2 also demonstrates – following the same procedure for $n = 3$, $b = 8$ one ends up with a negative bound on s. That happens because of the strong under-fluctuation in the background which lead us to exclusion of $s = 0$ at 95% C.L.

This issue was first addressed by Feldman and Cousins in [3], in a context of a more general problem of ambiguities in the selection of the confidence intervals (8.6) and (8.7). Authors of [3] consider a "flip-flopping physicist" – a highly hypothetical person who makes a decision on presentation of his experimental results after he did his measurements and had seen the results. Depending on the observed result he changes the way the x_1 and x_2 boundaries in (8.6) or (8.7) are chosen, breaking the very idea behind the construction of confidence belts.

Feldman and Cousins then go on by presenting a "unified" way of constructing confidence belts that behave smoothly around boundaries, allowing seamless transitions between different conventions for confidence interval construction, while preserving the frequentist nature of the resulting intervals.

Nevertheless even the Feldman-Cousins intervals have some unwelcome features that result in criticism and controversy. For example [4], such a "unified" construction for upper limit do depend on the expected number b of background events. Which leads to somewhat objectionable dependence of

the upper limit on b even when $n = 0$ and, obviously, $n_b = 0$ and the expected background should be irrelevant.

8.4 Bayesian intervals and Zech's interpretation

The complications discussed in previous chapter are all related to the fact that we do not fully exploit all our knowledge about the background (especially posterior) in the limit construction. An alternative way of dealing with that problem was proposed in [5, 6]. It is based on the Bayesian approach; application of Bayes's formula leads one to definition of the following limit on s with confidence $1 - \alpha$, obtained by:

$$\alpha = \frac{P(N \leq n|s+b)}{P(N \leq n|b)}. \tag{8.13}$$

The interval defined by (8.13) is called a "Bayesian credible interval" (with uniform prior). To avoid all the disputes, associated with the Bayesian approach, the frequentist interpretation of (8.13) was proposed by Zech in [7]. First note that we can see the random variable N in (8.11) as a sum of two independent random variables $N = N_s + N_b$. The probability distribution for N can then be presented as a convolution of independent distributions for numbers of signal and background events:

$$P(N = n|s,b) = \sum_{n_b=0}^{n} P(N_b = n_b|b)P(N_s = n - n_b|s)\,. \tag{8.14}$$

As it is stressed in [7], instead of using the original probability $P(N_b = n_b|b)$ of getting n_b background events, one should condition this probability to the already observed fact that n_b cannot be larger than n:

$$P(N_b = n|b, N_b \leq n) = P(N_b = n|b) \Big/ \sum_{n_b=0}^{n} P(N_b = n_b|b)\ . \tag{8.15}$$

The reason for such "renormalization" is that we are not expecting to get any extra information about the background process and not improving our knowledge about the background rate b. Combining (8.15) and (8.14) one gets:

$$P(N = n|s,b,N_b \leq n) = P(N = n|s,b) \Big/ \sum_{n_b=0}^{n} P(N_b = n_b|b)\ . \tag{8.16}$$

The corresponding likelihood function and the 95% confidence level upper bounds are shown on Figure. 8.3.

It so turns out, that the expression (8.16) coincides with the Bayesian expression. While there was a certain discussion about if it is Bayesian or frequentist approach, this "modified frequentist" CL_s technique is used as a standard de-facto approach in particle physics searches at CERN:

$$CL_s = \frac{CL_{s+b}}{CL_b}. \tag{8.17}$$

8.5 Hypothesis testing and nuisance parameters

To continue our discussion on presentation of experimental results of searches of new particle physics phenomena, we first will need to review some more basic concepts from statistics, related to *hypothesis testing*.

Statistic is a function of observables that is used to quantify certain properties of the observed data, usually interesting for the analysis in hand. And *test statistic* is a statistic that is useful for hypothesis testing – it is used to rate which of two given hypothesis is more favored by the data.

In particle physics experiments searching for new phenomena the *null hypothesis* H_0 is usually a statement that only known processes (backgrounds) are contributing to the experimental result. The "rival" to this hypothesis is a *signal hypothesis* H_1, which claims that there is an extra process contributing to observed data.

Neyman-Pearson lemma is applicable in case when H_1 and H_0 are two simple hypothesis – they should not involve any extra parameters. For example, one can say that H_1 states that expected event rate is $s + b$, while H_0 asserts that it is just b. The lemma then states that the best discrimination power for is achieved using likelihood-ratio test statistics:

$$Q = -2 \ln \frac{L(H_1)}{L(H_0)} = -2 \ln \frac{P(N = n|s + b)}{P(N = n|b)} \tag{8.18}$$

The Wilk's theorem [8] provides a very convenient property of the likelihood ratio test statistics. The distribution of Q in (8.18) asymptotically (with growing sample size) approaches the χ^2-distribution. This property makes it extremely easy to construct approximate confidence intervals for the test statistics.

Two important concerns complicates the simple picture of eq. (8.18). First of all we would prefer to present the results in term of a *signal strength modifier* μ which scales the signal cross section. Use of μ is preferable because we usually not only want to reject the signal hypothesis, but also quote the sensitivity of the experiment to weaker signals to account for possibilities of New Physics modifications, that might affect the signal expected cross-sections. The null hypothesis is, therefore, corresponds to $\mu = 0$, while H_1 and test statistics Q in (8.18) become dependent on μ:

$$Q_\mu = -2 \ln \frac{P(N = n|\mu \cdot s + b)}{P(N = n|b)}, \tag{8.19}$$

Second thing to be accounted for is the fact that, the experimental predictions are usually affected by systematic uncertainties – some details that we are not quite certain about, so it is necessary to introduce some degree of randomness, associated with this uncertainties. The standard way of dealing with systematic uncertainties is to introduce a number of the so-called *nuisance parameters* θ in addition to parameter μ that we are interested in determining from the experiment. The nuisance parameters describe possible deviations of the distributions of the expected experimental results due to systematic errors.

In the experiment we are not interested in determining these parameters and some procedure should be adopted to remove the nuisance parameters from consideration. A convenient way of removing nuisance parameter relies on the concept of the *profile likelihood* function:

$$\tilde{L}(\mu) = L(\mu, \hat{\hat{\theta}}(\mu)), \quad \hat{\hat{\theta}}(\mu) = \arg\max_\theta L(\mu, \theta). \tag{8.20}$$

Where $\hat{\hat{\theta}}(\mu)$ denotes values of nuisance parameters that maximize the joint likelihood for given value of θ. For the hypothesis testing instead of (8.18) the *profile likelihood ratio* is introduced:

$$\tilde{Q}_\mu = \begin{cases} -2 \ln \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(0, \hat{\hat{\theta}}(0))} & \hat{\mu} < 0, \\ -2 \ln \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})} & 0 \leq \hat{\mu} \leq \mu, \\ 0 & \hat{\mu} > \mu, \end{cases} \tag{8.21}$$

where numerator is the same profile likelihood as in (8.20), while in the denominator the maximum likelihood estimators for both μ and nuisance pa-

rameters θ are substituted, with $\hat{\mu}$ being constrained to the $[0, \mu]$ range in order to provide only upper bounds on μ.

The choice of the test statistics (8.21) is motivated by the asymptotic formulae, obtained in [9]. It turns out that, similarly to the Wilk's theorem, the asymptotic behavior of $\tilde{Q}_\mu$ follows a *non-central* χ^2 distribution:

$$P(\tilde{Q}_\mu = \tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \leq \frac{\mu^2}{\sigma^2}, \\ \frac{\sigma}{2\mu\sqrt{2\pi}}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu+\mu^2/\sigma^2)^2}{4\mu^2/\sigma^2}\right] & \tilde{q}_\mu \leq \frac{\mu^2}{\sigma^2}, \end{cases} \tag{8.22}$$

where the variance σ can be easily evaluated using the so-called "Asimov data set" – Value of the test statistics (8.21) is evaluated for the background-only "observation" with the observed event count being equal to the expected background rate and all nuisance parameters θ are set to their nominal values $\tilde{\theta}$. The resulting test statistics for Asimov data set is related to the variance of the distribution (8.22) simply as:

$$\sigma^2 = \frac{\mu^2}{q_{\mu,A}}. \tag{8.23}$$

As a consequence, one can obtain the expected limits and estimate the sensitivity of the experiment without performing a CPU-consuming procedure Monte Carlo generation of pseudo-experiments to estimate the distribution of the test statistics.

8.6 CL_s technique for search results presentation.

We complete this chapter by summarizing the entire procedure for limit construction that is used for results presentation of analyses presented in this work.

In all these analyses, the distribution in invariant mass of Higgs particle is examined. Each bin of the distribution is considered as a separate channel. The likelihood is, therefore, constructed as product of Poisson distributions for each bin, times the *p.d.f* for the nuisance parameters:

$$\begin{aligned} L(\mu, \theta) &= p(\tilde{\theta}; \theta) \cdot \prod_{i=1} Poisson(n_i;\ \mu \cdot s_i(\theta) + b_i(\theta)) \\ &= p(\tilde{\theta}; \theta) \cdot \prod_{i=1} \frac{(\mu \cdot s_i(\theta) + b_i(\theta))^{n_i}}{n_i!} e^{-\mu \cdot s_i(\theta) - b_i(\theta)} \end{aligned} \tag{8.24}$$

The nuisance parameters θ are assumed to parametrize the values of the signal $s_i(\theta)$ and background $b_i(\theta)$ expected rates. The sources of the uncertainties are expected to be either uncorrelated or 100% correlated.

The parameters $\tilde{\theta}$ in (8.24) describe the distribution of the nuisance parameters, setting the "default" value of θ. Note that probability distribution $p(\tilde{\theta};\theta)$ does not consider θ as a random variable – the proper *Bayesian* approach would be to consider a systematic error *p.d.f*s $\rho(\theta;\tilde{\theta})$. Yet we are, following [10], re-interpreting such distributions in frequentist language by means of Bayes' theorem:

$$\rho(\theta;\tilde{\theta}) \sim p(\tilde{\theta};\theta) \cdot \pi_\theta(\theta), \tag{8.25}$$

with $\pi_\theta(\theta)$ being hyper-priors, that is prior distribution for a parameter of a prior distribution. The three types of *p.d.f*s are considered for distributions of nuisance parameters: normal, log-normal and gamma distributions. It turns out (see e.g. [10]) that the three distributions are trivially transformed when the hyper-prior is kept flat – the normal, log-normal and gamma distributions on θ gets transformed into the same the normal, log-normal and gamma distributions on $\tilde{\theta}$.

The likelihood function (8.24) is then used to construct the profile likelihood ratio test statistics (8.21) and the observed value $\tilde{q}_\mu^{obs}$ of the $\tilde{Q}_\mu$ statistics is computed based on the observed data. After that we proceed by computing the CL_s upper bound (8.17) for a given μ:

$$CL_s(\mu) = \frac{P\left(\tilde{Q}_\mu \geq \tilde{q}_\mu^{obs} | signal + background\right)}{P\left(\tilde{Q}_\mu \geq \tilde{q}_\mu^{obs} | background\right)}. \tag{8.26}$$

After that one quotes the value of $\mu = \mu^{95\%CL}$, called "the 95% Confidence Level upper-limit on μ", which is obtained by adjusting μ until $CL_s = 0.05$. Together with the $\mu^{95\%CL}$ value the expected limits for the null hypothesis are presented in from of a $\pm 1\sigma$ and $\pm 2\sigma$ bands by plotting the values of μ at which corresponding *c.d.f* for background-only hypothesis crosses the values of 2.5%, 16%, 50%, 84%, and 97.5%.

Figure 8.4 shows an example of the exclusion plot for Higgs searches. This is an actual plot for the $H \to ZZ \to 4\ell$ process from the work [11]. The black dots on the plot are showing the values of $\mu^{95\%CL}$ for various m_H

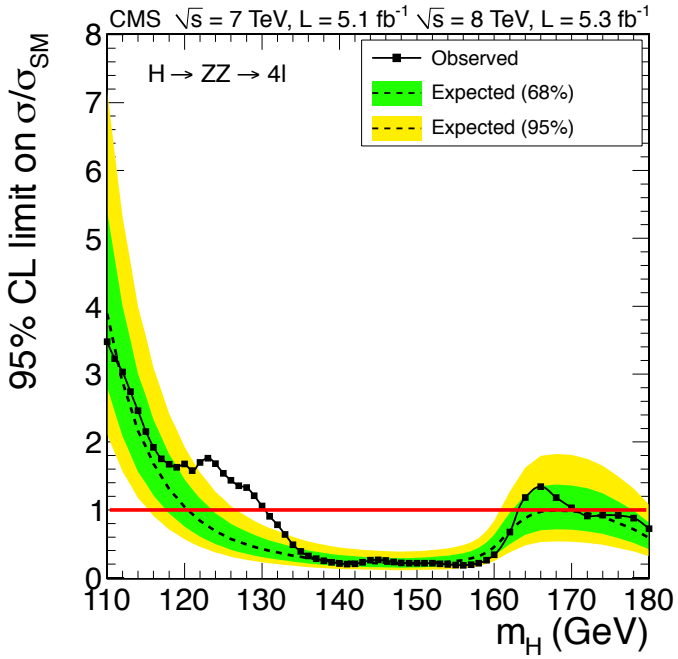


Figure 8.4: Example of the CL_s exclusion plot (from [11]), showing observed 95% CL_s limit on the SM Higgs boson signal strength together with 68% ($\pm 1\sigma$) and 95% ($\pm 2\sigma$) expected limits band for background-only hypothesis.

hypotheses. The green and yellow bands are demonstrating the background-only band for expected limits. One can see that this analysis reaches the 95% CL exclusion sensitivity (where expected limit crosses $\mu = 1$ line) for Higgs mass hypothesis in the range $m_H \in [120, 180]$ and that the SM Higgs hypothesis is excluded in mass ranges $m_H \in [130, 164] \cup [170, 180]$.

Bibliography

[1] A. N. Kolmogorov. "Basic notions of the probability theory" (1975).

[2] J. Neyman. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability". *Royal Society of London Philosophical Transactions Series A* 236 (Aug. 1937), pp. 333–380.

[3] Gary J. Feldman and Robert D. Cousins. "A Unified approach to the classical statistical analysis of small signals". *Phys.Rev.* D57 (1998), pp. 3873–3889. arXiv:physics/9711021 [physics.data-an].

[4] G. Cowan. "Statistics". *Phys.Lett.* B667 (2008), pp. 320–329.

[5] O. Helene. "Upper Limit of Peak Area". *Nucl.Instrum.Meth.* 212 (1983), p. 319.

[6] O. Helene. "Determination of the upper limit of a peak area". *Nucl.Instrum.Meth.* A300 (1991), pp. 132–136.

[7] G. Zech. "Upper limits in experiments with background or measurement errors". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 277.23 (1989), pp. 608 –610. ISSN: 0168-9002.

[8] S.S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". *Annals Math.Statist.* 9.1 (1938), pp. 60–62.

[9] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". *Eur.Phys.J.* C71 (2011), p. 1554. arXiv:1007.1727 [physics.data-an].

[10] ATLAS Collaboration. "Procedure for the LHC Higgs boson search combination in summer 2011" (2011).

[11] CMS Collaboration. "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$ = 7 and 8 TeV". *JHEP* 1306 (2013), p. 081. arXiv:1303.4571 [hep-ex].

Chapter 9

Search for SUSY Higgs boson decaying into a pair of b-quarks

In Chapter. 4.2 we have discussed that the design of the LHC machine together with ATLAS and CMS detectors is targeted at the detection of an electroweak resonance with a clear *full-leptonic* signature. That was done primarily due to the fact that detection of hadronic signatures (even with increased cross-sections and branching ratios) will be greatly complicated by huge QCD backgrounds and by strong reduction of experiment resolution due to jets in the final state.

Some BSM extensions might provide a way out of these complications – the extra particles and/or interactions can either result in enhanced cross-sections or with an easily distinguishable event signature.

Analysis in this chapter was performed using 2011 data and targets MSSM extension of the Standard Model (see Section 3.4). In the MSSM large values

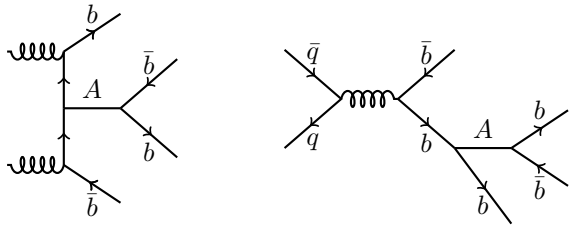


Figure 9.1: Feynman diagrams for the process under investigation.

of the $\tan\beta$ parameter result in enhanced Yukawa couplings of fermions to the scalars of the MSSM Higgs sector. If one targets the events where the scalar is produced in association with a pair of b-quarks and decays into a $b\bar{b}$ pair (see Figure 9.1) then he can simultaneously benefits from the $\tan\beta$ enhancement and from multiple b-tagging signature in the final state.

The particular benchmark point for such search, called "m_h^{max} scenario" was suggested in [1, 2]. Motivation for the benchmark scenario and corresponding set of MSSM parameters are briefly reviewed in the Section 3.4.3. Extra enhancement for the processes on Figure 9.1 is expected in this scenario, since all the three sypersymmetric scalars are nearly degenerate in that case, so they contribute independently to the total signal cross-section.

The p_T and $|\eta|$ spectra of the four b-quarks in signal events for $m_A = 120$ GeV are presented on Figure 9.2. As one can see, the two leading b-quarks usually originate from the Higgs decay. They are also concentrated in the central region $|\eta| < 2.5$, where b-tagging algorithms can be effectively used.

The two b-quarks that are not produced by the scalar decay (the associated b-quarks) have a considerably softer p_T spectrum. The fourth leading b-quark is usually very forward. That makes the corresponding fourth jet very difficult to reconstruct and even more difficult to b-tag.

All in all – the generic search strategy is to collect a sample with *three* b-tagged jets, and consider the invariant mass of the *two leading* b-tagged jets. The observed distribution in the reconstructed invariant mass is then tested against the background hypothesis.

The dominant background for the process is the QCD multijet events. The most complex feature of the analysis is the data-driven determination of the QCD background shape – one cannot rely on the Monte Carlo for the shape and normalization prediction because of prohibitive computational demands. The QCD multijet background prediction was extracted from data using two conceptually different approaches: the matrix method and, for cross-checking and evaluation of systematic uncertainties, the "Hyperball" method.

9.1 Online data selection

Collection of all the events with at least three jets is hopeless due to enormous background from multi-jet QCD events. The event rate must be first reduced

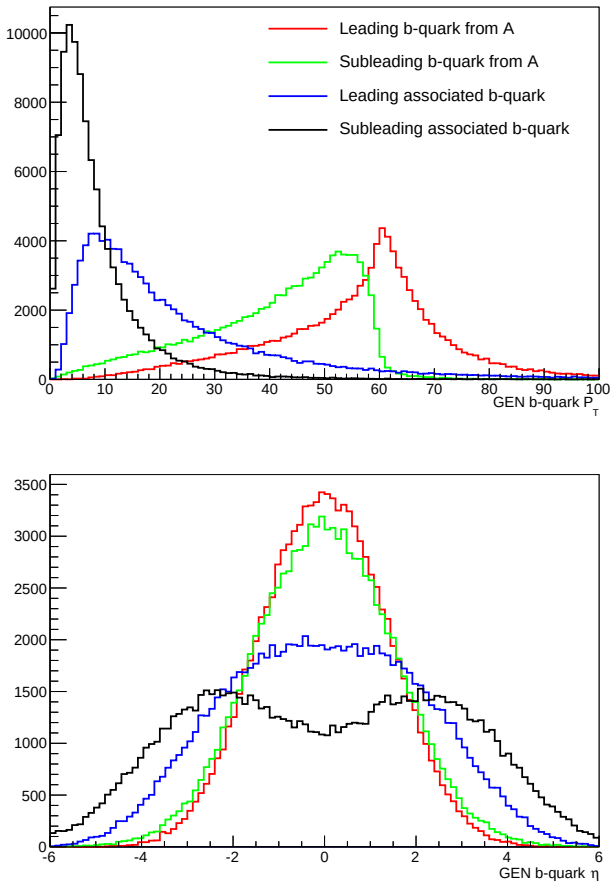


Figure 9.2: The p_T and η distributions of the four b-quarks of in the $m_A = 120$ GeV signal events.

HLT paths *(L1 seed)*	run range	triggered events	$\int \mathcal{L} dt$ [pb^{-1}]
HLT_Mu12_CentralJet30_BtagIP	163738-165633	3 027 717	180.9
L1_SingleMu7			
HLT_Mu12_DiCentralJet30_BtagIP3D	165970-166967	4 532 555	537.1
L1_SingleMu10			
HLT_Mu12_DiCentralJet20_DiBtagIP3D1stTrack	167039-173198	2 244 550	1108.6
L1_SingleMu10			
HLT_Mu12_eta2p1_DiCentralJet20_DiBtagIP3D1stTrack	173236-176469	1 237 147	652.2
L1_Mu10_Eta2p1_DoubleJet_16_8			
HLT_Mu12_eta2p1_DiCentralJet20_DiBtagIP3D1stTrack*	176545-180252	5 690 304	2326.8
*L1_Mu10_Eta2p1_DoubleJet_16_8**			
All		16 732 273	4805.7

Table 9.1: HLT paths used through different data taking periods.

HLT Path	signal efficiency [%]
HLT_Mu12_CentralJet30_BtagIP	3.64 ± 0.19
HLT_Mu12_DiCentralJet30_BtagIP3D	2.28 ± 0.11
HLT_Mu12_DiCentralJet20_DiBtagIP3D1stTrack	1.66 ± 0.07
HLT_Mu12_eta2p1_DiCentralJet20_DiBtagIP3D1stTrack	1.65 ± 0.06

Table 9.2: Estimated signal ($m_H = 120$ GeV) efficiences for the considered HLT paths.

by online selection at trigger level. The online selection strategy, however must retain a comparably high signal efficiency. A number of combined *muon + jet + b-tag* triggers was employed to acieve these goals.

The data taking strategy throughout year 2011 followed a gradual increase in the instantaneous luminosity (see Section 4.2.2). Corresponding adjustments of the HLT triggering strategy was necessary to maintain both acceptable selection rate and high signal efficiency. As a result, four HLT paths, tweaked to have acceptable rates, were sequentially used in this analysis.

The path design always contained a reasonably high-p_T muon without isolation requirements. The muon is used to distinguish events with semileptonically decaying b-quarks in the final state. The $b \to \mu$ branching ratio is about 10%, which substantially reduces signal efficiency, but that is compensated by overall reduction of non b-quark QCD background.

The triggers require the presence of energetic central jets and apply an on-line b-tagging based on the THCP algorithm, as described in Chapter. 7.3. Brief description of the different trigger paths used in the online data selection follows:

HLT_Mu12_CentralJet30_BTagIP:
The trigger requires sequentially $p_T > 7$ GeV muon at L1, $p_T > 10$ GeV at L2 and $p_T > 12$ GeV at L3 with increasing quality requirements. Simultaneously, the presence of $E_T > 30$ GeV and $|\eta| < 2.6$ (central) jet is demanded. `BTagIP` (see Section 7.3) online b-tagging implementation was used in this HLT path.

HLT_Mu12_DiCentralJet30_BTagIP3D:
Because of the increase of the instantaneous luminosity, the $p_T > 10$ GeV muon at L1 was used. Two central jets with $E_T > 30$ GeV are required. The online b-tagging is improved by using 3D primary vertex reconstruction.

HLT_Mu12_DiCentralJet20_DiBTagIP3D1stTrack:
Here, the threshold on the two leading central jets was "traded" for an extra online b-tag. The requirement on the transverse momentum of the two jets at L2 is reduced to 20 GeV. In order to maintain a low rate, the requirement on number of b-tagged jets is increased to two.

Another modification is in the online THCP algorithm. S_{IP} cut is applied to the first track instead of the second track.

HLT_Mu12_eta2p1_DiCentralJet20_DiBTagIP3D1stTrack:
With the increase of the instantaneous luminocity, switching to the L1 bit `L1_Mu10_Eta2p1_DoubleJet_16_8_Central` was done. It requires at least two central $|\eta|$ $\leq$2.1 jets with $p_T \geq 16,8$ GeV and a muon with$p_T \geq 10$ GeV.

HLT_Mu12_eta2p1_DiCentralJet20_DiBTagIP3D1stTrack*:
The path is virtualy the same as the previous one, but with updated set of online jet energy corrections that increase the performance of the jet E_T efficiency curves.

Table. 9.1 summarizes the different HLT trigger paths and their corresponding L1 trigger seeds used in the analysis for various data taking periods.

9.2 Offline data selection

The offline data selection serves two purposes – it partially reduces the QCD background by requiring a number of cuts and using CSV b-tagging algorithm. QCD background cannot be completely removed by the selection cut. Development of a data-driven method for the QCD background prediction is necessary – the offline data selection also serves as a first step in preparation of a sample for such prediction.

9.2.1 Physics objects

Details of the physics objects reconstruction, their indentification and necessary corrections are presented in the Chapter 6.

Muons used in the analysis are `Global Muons` with tight ID selection criteria for the 2011 period – see Table 6.2.

Jets are standard PF-reconstructed jets with anti-k_t algorithm, $\Delta R = 0.5$. Loose JetID criteria (see Table 6.1) together with all the Jet Energy Corrections, provided by the corresponding POG were also applied.

9.2.2 Event selection

The offline pre-selection cuts on the reconstructed objects are:

- at least three jets should be present in the event. For the jets to be b-tagged it is necessary to consider only those within the tracker acceptance – that motivates the pseudorapidity cut of $|\eta| < 2.6$.
- transverse energies of the two of these jets should satisfy $E_T^1, E_T^2 > 30$ GeV and for the third jet $E_T^3 > 20$ GeV
- to avoid any ambiguities in b-tagging, the ΔR distance between each pair of jets is constrained to $\Delta R_{ij} > 1$
- first two jets, ordered in E_T, are then required to have CSV b-tag discriminator value of at least $CSV_{1,2} > 0.8$
- the global muon with $p_t > 15$ GeV must also be present and should belong to one of the two leading jets.

9.2.3 Combining HLT paths

In our analysis we use several HLT paths, while we want to analyze the data altogether regardless of a trigger path. This can be a source of a bias, as it is demonstrated on Figure 9.3,

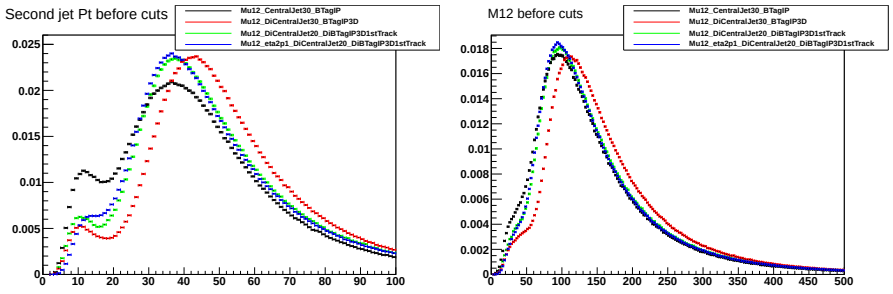


Figure 9.3: Offline sub-leading jet p_T (left) and m_{12} (right) distributions in events triggered by different HLT paths used in the analysis.

where the distributions in the p_T of the second jet and in the invariant mass of the first and second jets are shown. All the histograms were rescaled

to have the same integral value; colors on these plots represent different HLT paths:

Black `HLT_Mu12_CentralJet30_BtagIP`;

Red `HLT_Mu12_DiCentralJet30_BtagIP3D`;

Green `HLT_Mu12_DiCentralJet20_DiBtagIP3D1stTrack`;

Blue `HLT_Mu12_eta2p1_DiCentralJet20_DiBtagIP3D1stTrack`.

As one can see, there is a strong difference between the spectra for these HLT paths, which is expected since we have different L1 requirements on the sub-leading jets and different kinematic bounds on both jets and on the triggering muon.

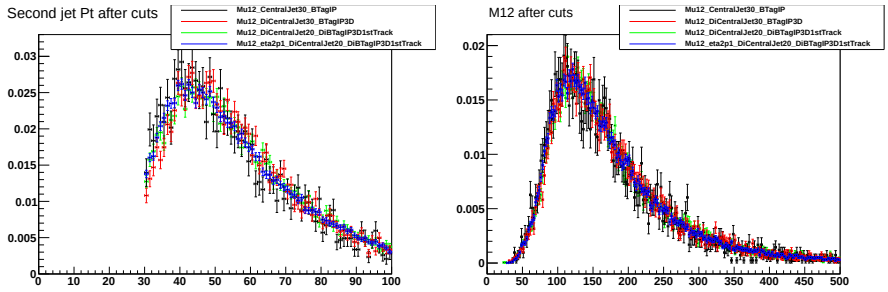


Figure 9.4: Offline sub-leading jet p_T (left) and m_{12} (right) distributions in events triggered by different HLT paths. The coloring is the same as on Figure 9.3

The same (properly rescaled) distributions after full analysis cuts are presented on Figure 9.4. The distributions are now almost identical – the difference between them, given the reduced event count after the cuts, can be attributed to statistical fluctuations.

Similar comparison was performed for other kinematic parameters (like η_i, ϕ_i, ΔR_{ij}, e.t.c.), yielding the same results. Which allows us to conclude that the bias, associated with "mixing together" all the HLT paths is eliminated by our offline cuts. So we can analyze the full statistics at once, regardless of the trigger path.

Cut	Mu12 CentralJet30 BTagIP	Mu12 DiCentralJet30 BTagIP3D	Mu12 DiCentralJet20 DiBTagIP3D1stTrack	Mu12_eta2p1 DiCentralJet20 DiBTagIP3D1stTrack	Mu12_eta2p1 DiCentralJet20_NewJEC DiBTagIP3D1stTrack	All
All	3027717	4532555	2244550	1237147	5690304	16732273
$P_T^{\mu} > 15$ GeV	1757902	2678935	1337394	742231	3222677	9739139
$\#jets >= 3$	665962	1245655	639616	404082	1556012	4511327
$\Delta R_{ij} >= 1$	513981	957884	498996	315284	1219439	3505584
$CSV(1^{st} - jet) > 0.8$	242982	492734	297838	184075	714506	1932135
$CSV(2^{nd} - jet) > 0.8$	52345	112428	162029	99175	387708	813685
μ in 1^{st} or 2^{nd} jet	50708	108551	156147	95760	374774	785940
$CSV(3^{rd} - jet) > 0.7$	3245	7323	12796	7623	29208	60195
$\int \mathcal{L} dt\ pb^{-1}$	180.9	537.1	1108.6	652.2	2326.816	4805.7

Table 9.3: Data reduction after each selection cut for the various trigger paths

9.3 Data-driven multi-jet background determination

As it was mentioned before, the primary background for the studied process is coming from multi-jet production from hard QCD scattering processes. The Monte-Carlo simulation of such processes is not suitable because of lack of the corresponding NLO/NNLO calculations as well as due to the uncertainties of b-jet production in the events with gluon-splitting.

Two data-driven methods were developed and used to evaluate the multi-jet background. The first method is the matrix method where signal-poor data sample (the "control region") is selected and the shape of the background (together with its normalization) are extrapolated to the signal region.

The second method is the so-called "Hyperball" method – a specific variant of the "nearest neighbor" multivariate method. It introduces the "distance" between events in parameter space and infers the probability of an event to have tree b-tags by averaging over the "neighbors" in the training sample.

9.3.1 Matrix method

The central idea of the matrix method of data-driven background estimation can be summarized as follows: The 2-b-tagged (bbj) sample is the *data* sample with only two first leading jets b-tagged and no tagging requirement for the third jet. The bbj sample contains very small fraction of the signal events. Let us denote the distribution of some event variable x in the bbj sample as $F(x; bbj)$. The distribution of x in the 3-b-tag sample (bbb) can be obtained by multiplication by the probability of the third jet to be tagged:

$$F(x; bbb) = F(x; bbj) \times P(\text{3rd jet b-tagged}). \tag{9.1}$$

The probability $P(\text{3rd jet b-tagged})$ depends on various properties of the third jet itself and on the remaining characteristics of the event. We further split that probability into several flavor-dependent parts:

$$P(\text{3rd jet b-tagged}) = \epsilon_b f_b + \epsilon_c f_c + \epsilon_{uds} f_{uds}, \tag{9.2}$$

where ϵ_q is the probability of the jet with flavour q to be b-tagged and f_q is the fraction of the q-quarks that will produce the third jet in the sample. The

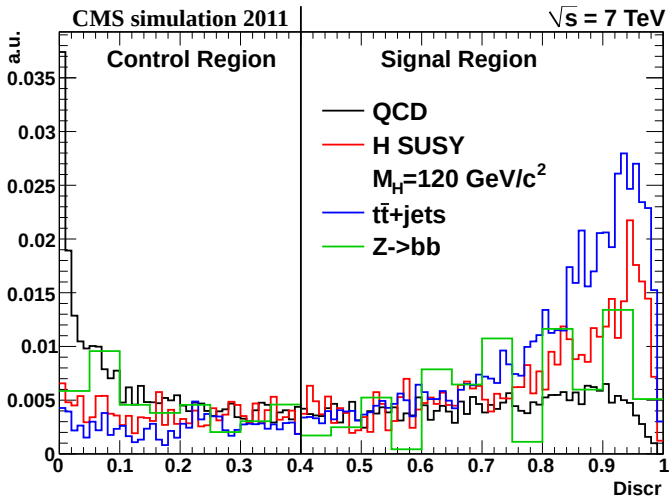


Figure 9.5: The likelihood discriminant used for the selection of control region.

b-tagging efficiencies ϵ_q are well-known – their determination as functions $\epsilon_q(E_{T3}, |\eta)3|, N_{ch3})$ was performed in [3] together with a data-driven cross-check. Data/MC scale factor of 0.95 ± 0.04, reported in [3] was also applied by us.

Flavor fractions f_q, on the other hand, are process-specific and their determination for the current analysis is much more involved. Firstly, using a multivariate likelihood discriminator *control region* with low signal-to-background ratio is defined. The variables used for the discriminator are:

$$\begin{array}{llll} p_{Ti}, & \Delta\phi_{ij}, & \Delta\eta_{12}, & min(\Delta\eta_{ij}), \quad max(\Delta\eta_{ij}), \\ \Delta R_{12}, & \Delta\phi_{H3}, & \Delta\eta_{H3}, & min(\Delta R_{ij}), \\ \alpha_{H3}, & \cos\theta^*, & N_{jets}, & \end{array} \tag{9.3}$$

where $(i, j = 1, 2, 3)$ are the p_T-ordered jet numbers, H is the combination of first and second jets, α is the angle between two vectors and θ^* is the angle between H flight direction and the direction of the first jet in the H rest frame.

Using these variables, two likelihood discriminators were trained – one for the Higgs hypothesis in the low mass region: $m_H = 120, 130, 140$, and another one for the high-mass region: $m_H = 250, 300, 350$. The distributions for

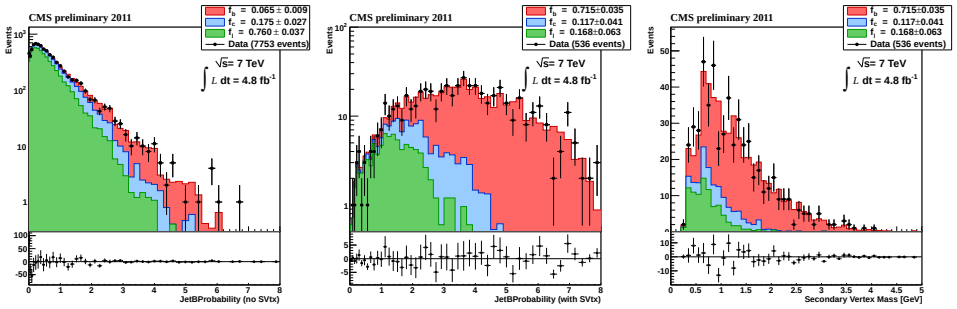


Figure 9.6: Example of the flavor fractions extraction for one bin of the $f_1\left(E_{T3}, |\eta_3|\right)$ function.

signal and various backgrounds for the low-mass discriminator are shown on Figure 9.5.

It must be stressed that this discriminator was not used for selection of neither signal nor background – its sole purpose was to determine the region of phase space where multijet background dominates the signal contribution. Such a control region is constructed for data-driven determination of flavor fractions f_q in (9.2).

One proceeds by parametrising the flavor fractions f_q using two factors. The first factor depends on E_{T3} and $|\eta_3|$ (as ϵ_q does) the second factor incorporates information about event topology through ΔR_{12} and ΔR_{H3} distances:

$$f_q = f_1\left(E_{T3}, |\eta_3|\right) \times f_2\left(\Delta R_{12}, \Delta R_{H3}\right). \tag{9.4}$$

The normalization information is assumed to be contained only in f_1. The f_2 factor is used for shape adjustment and normalization to unity.

For each bin in $E_{T3} \otimes |\eta_3|$ and in $\Delta R_{12} \otimes \Delta R_{H3}$ planes we performed the template fit of two b-flavor discriminating variables: `JetBProbability` (see Chap. 7.2) and `TagMass` – the invariant mass, associated with the reconstructed secondary vertex (if vertex is not reconstructed, then only JBP is used). The fitted templates are obtained from QCD Monte-Carlo simulation for individual quark flavors. Figure 9.6 shows an example of the f_q extraction for one bin of the $f_1\left(E_{T3}, |\eta_3|\right)$ distribution.

Flavor fractions f_q and b-tagging efficiencies ϵ_q can then be used in (9.1) and (9.2) to obtain the distribution of any variable of interest for the *bbb* sample. Figure 9.7 shows the distribution in invariant mass of the two leading

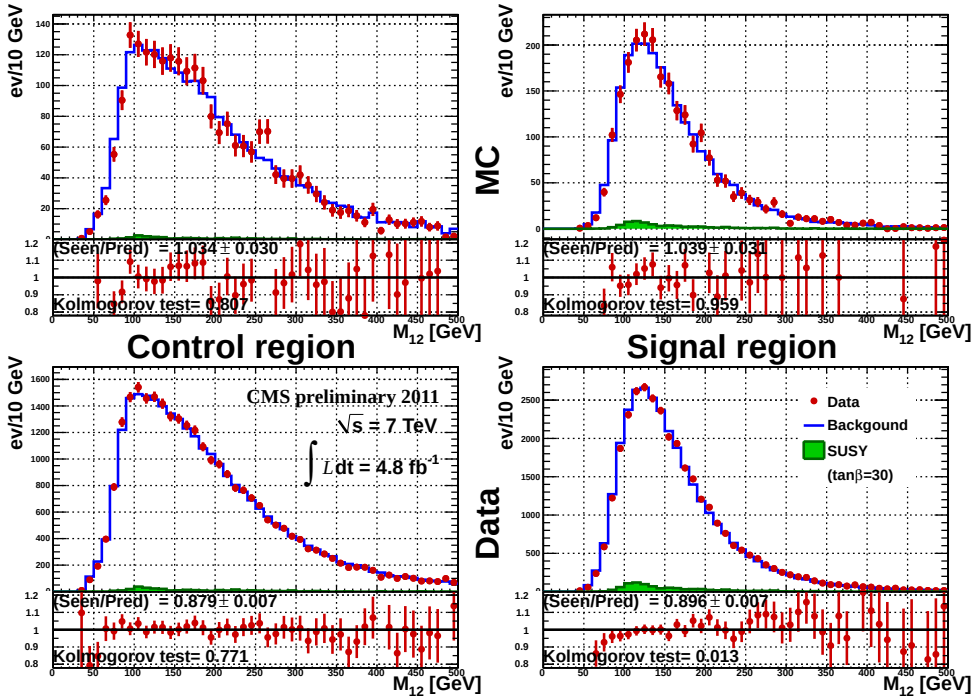


Figure 9.7: Distribution of the invariant mass of the two leading jets, illustrating the matrix method background determination.

jets m_{12} for bbb events, both predicted (blue curve) and actual ones (red dots). Top row shows MC distributions for control and signal regions, while bottom row demonstrates these for data.

The predicted background distributions was normalized to data – the normalization factor is also shown on Fig. 9.7 denoted as "Seen/Pred". As one can see, normalization and shape agree well with the Monte Carlo. For the data, shape is also well-reproduced, while normalization is noticeably (about 10%) overestimated. The bias most probably arises from biases in the template fitting procedure.

Similar distributions are produced for a number of other variables: E_T and $|\eta|$ of the jets and value of the likelihood discriminator. The predicted background was also compared for different HLT paths. For all the distributions the shape was very well reproduced, while the normalization required roughly the same scaling factor for each distribution both for control and sig-

nal regions.

The above observations allow us to conclude that it is possible to extract the normalization from the control region and use it for prediction of the background in the signal region.

9.3.2 Hyperball method

The alternative data-driven technique for background estimation uses the so-called "Hyperball" [4] algorithm – the specific variant of a general "nearest neighbor" class of multivariate algorithms. The basic idea behind the algorithm is the estimation of a certain function on the space of event parameters by averaging over several "nearest" events in the parameter space. These neighbor events are determined by introducing a "distance" measure for a pair of events evaluated on their parameters.

Previously, Hyperball was used in high-energy physics in [5] for improvement of di-b-jet resolution. The function of interest was the invariant mass bias and it was estimated from Monte Carlo events by means of the Hyperball method. Correcting for the bias significantly improved the mass resolution.

In this analysis the estimated function is the probability of an event to have three b-tags. The training sample consists of the data sample containing at least three jets, passing kinematic cuts from Section 9.2. The "bjj sample" is then constructed by requiring only the first jet to be b-tagged with CSV algorithm and by applying the "control region" selection by cutting the likelihood discriminant $LD < 0.4$ as it is described in the previous section.

The distribution of any variable for the events with 3 b-tagged jets (bbb events) is obtained by reweighting the distribution of that variable in bjj sample. Weights are calculated by picking N_H events in the training sample that are closest to the test event in the multidimensional parameter space. Weight is then equals to the weighted fraction of bbb events among the chosen N_H:

$$\text{weight} = \frac{\sum\limits_{bbb \in N_H} D_{bbb}^{-2}}{\sum\limits_{bjj \in N_H} D_{bjj}^{-2}}. \tag{9.5}$$

The "closeness" measure D mentioned above is defined on the set of variables describing an event. The distance between one event x_i (the test event)

and another event y_i (the training event) is then defined as a weighted euclidean distance D:

$$D^2 = w_i^2(x_i - y_i)^2. \tag{9.6}$$

The weights w_i^2 are themselves functions of x_i – for a given test event they represent how sensitive is the fraction of the *bbb* events to changes in i-th variable at that point in parameter space.

In practice, the weights are computed as numerical derivatives of dependence of the probability on each of the variables. Each variable range was split into a number of non-uniform bins. The fraction of the *bbb* events is interpolated between bins and the derivative of that dependece is used as w_i in (9.6).

When the test event lies close to its threshold (like p_T or $|\eta|$ cutoff) the hyperellipsoid of "closest" events is not centered with respect to test event. This introduces a bias in the *bbb*-probability estimation. The bias was removed, using linear interpolation over the events that are close to the same threshold in the "problematic" variable.

The algorithm was trained with 400000 events using the 12 variables listed below:

$$\begin{array}{cccc} E_{T1} & E_{T2} & E_{T3} & p_{T1}+p_{T2} \\ \eta_1 & \eta_2 & \eta_3 & p_{T2}+p_{T3} \\ \Delta\phi_{12} & \Delta R_{12} & M_{12} & p_{T1}+p_{T2}+p_{T3} \end{array} \tag{9.7}$$

The choice of the number N_H of the "neighboring" events used in the averaging (9.5) is delicate problem. On one hand it is constrained by the size of the sample and by the number of dimensions of the sample – too large N_T will average over huge subspace of events. On the other hand, too small N_T will introduce a statistical uncertainty in the *bbb*-probability estimate. The balance between the two requirements is achieved at values of about $N_H \sim 100$.

Figure 9.8 shows the comparison between predicted and actual *bbb* distributions for QCD Monte Carlo samples for low-mass and high-mass regions. As on Fig. 9.7 the prediction is shown as the blue histogram, while the red dots show the actual *bbb* distribution. Again, one can see that the shape agrees very well both for control and signal regions.

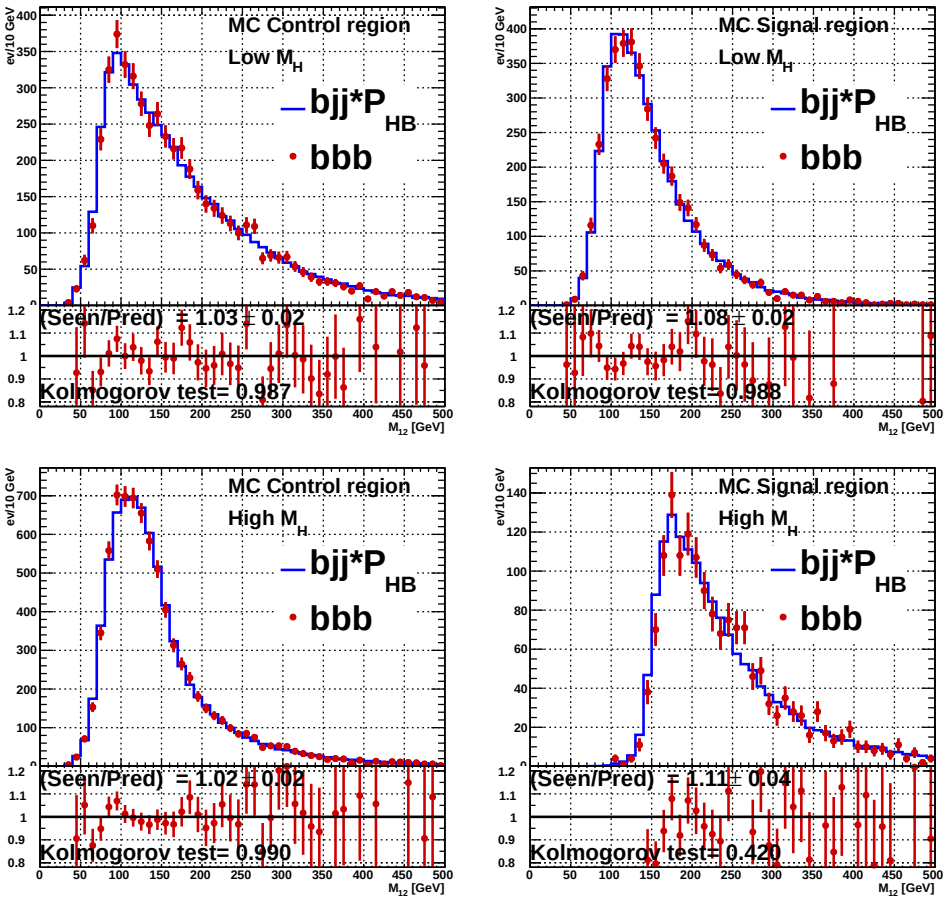


Figure 9.8: Actual and predicted invariant mass distribution by the "hyper-ball" method. For low (top row) and high-mass (bottom row) ranges of QCD simulated events.

9.4 Systematic uncertainties

In the discussion above we've seen that there is a number of *systematic effects* – certain biases associated to imperfections of simulations, selection efficiencies, energy and momentum resolutions, e.t.c. Such effects can be estimated and corrected for, but these corrections will have some uncertainty, associated to such corrections [6].

Sytematic error or *systematic uncertainty* is the uncertainty both in such corrections and in other parameters that our analysis depends on. The effect of the systematic uncertainties must be evaluated for a given analysis and accounted for in the results presentaiton using nuisance parameters (see Section 8.5).

In the following, the evaluation of the Parton Distribution Function (PDF) systematic uncertainty will be discussed in detail as an example. Then more brief overview of the effects of various systematic uncertainties of this analysis will be presented.

9.4.1 Example of evaluation of PDF systematic uncertainty.

Parton Distribution Functions describe momentum distribution of strongly interacting particles – quarks and gluons – inside the proton. By convention they re defined as functions of the momentum transfer Q^2 and the Bjorken variable x:

$$\sigma = \sum_{f,g} \int dx_1 dx_2 \; F_f(x_1, Q^2|, \mathbf{S}) F_g(x_2, Q^2|\mathbf{S}) \; \sigma_{f,g}(x_1, x_2, Q^2), \qquad (9.8)$$

where f and g are flavors (we will not mention them further) and $\mathbf{S}$ – is the set of N parameters for the PDF.

PDF parameters $\mathbf{S}_i$ are chosen in such a way as to correspond to eigenvectors of covariance matrix. Each parameter in the set is then varied up and down within tolerance, obtaining $2N$ "shifted" parameter sets $\mathbf{S}_\mathbf{i}^\pm$ $(i = 1 \ldots N)$.

The "brute-force" method of estimating PDF uncertainties would be just to repeat the MC for each parameter set $\mathbf{S}_\mathbf{i}^\pm$, obtaining shifted values for interesting observables. Given the typical value of $N \simeq 20$, it would take a lot of effort to do.

PDF weights

Since PDFs are just multiplicative factors, one can take a more efficient approach; namely, for each event a number of "PDF weights" W is calculated:

$$W_i^{\pm} = \frac{F(x_1, Q^2|\mathbf{S_i^{\pm}})F(x_2, Q^2|\mathbf{S_i^{\pm}})}{F(x_1, Q^2|\mathbf{S})F(x_2, Q^2|\mathbf{S})}$$

The observables are then weighted event-by-event with these $W_i^{\pm}$. It is important to note that weights can be calculated after the "main" MC, based only on kinematic parameters $x_{1,2}$, Q^2 and original PDF values $F(x_{1,2}, Q^2)$ (all the necessary data is usually stored in the CMSSW `GenEventInfoProduct` collection).

PDF sources

There are number of different PDF sources. We used the ones that are suggested in PDF4LHC recommendations [7, 8] for estimation of the PDF uncertainties. For NLO these are CTEQ66 [9], MSTW2008nlo [10] and NNPDF20 [11]. For NNLO it is just MSTW2008nnlo.

Uncertainty on rate

The basic observable that is interesting for us is the amount of events, passing the analysis cuts. The weighted version of this observable is obtained trivially by summing weights for each events (instead of just counting events).

The results for $M_H = 120\,GeV$:

$M_H = 120\,GeV$, 1093650 events		
PDF source	Central rate [events],[%]	Uncertainty in rate [+%/-%]
The Original PDF	2891, 100.0%	
CTEQ66	2942, 101.7%	+2.548 / -2.714
MSTW2008nlo	3078, 106.5%	+2.411 / -2.861
NNPDF20	3236, 111.6%	+2.815 / -2.773
MSTW2008nnlo	3085, 106.7%	+2.261 / -2.489

And for $M_H = 250\,GeV$:

$M_H = 250\,GeV$, 550000 events		
PDF source	Central rate [events],[%]	Uncertainty in rate [+%/-%]
The Original PDF	5506, 100.0%	
CTEQ66	5106, 92.7%	+4.702 / -4.390
MSTW2008nlo	5317, 96.6%	+3.217 / -4.036
NNPDF20	5537, 100.6%	+4.337 / -3.974
MSTW2008nnlo	5311, 96.5%	+3.325 / -3.412

A little explanation on what is a "central rate": the PDFs that are used in the MC generation are not necessarily the same as the PDFs in the estimation of systematics. So the relative change of the new PDF is calculated, based on known $x_{1,2}$ and Q^2.

$$W_{\text{central}} = \frac{F_{\text{new}}(x_1, Q^2|\mathbf{S}_{\text{new}})F_{\text{new}}(x_2, Q^2|\mathbf{S}_{\text{new}})}{F_{\text{original}}(x_1, Q^2|\mathbf{S}_{\text{original}})F_{\text{original}}(x_2, Q^2|\mathbf{S}_{\text{original}})}$$

Where $\mathbf{S}_{\text{new/original}}$ – are both unshifted ("the best") parameters for a given PDF.

Interpretation of high deviations in central rate

In our MC, the CTEQ6L1 [9] (which is LO fit with LO $alpha_s$) was used, while we set actual cross-section according to FeynHiggs [12] calculations, which are NLO. Therefore are just taking the uncertainties with respect to CTEQ66. Namely:

$$M = 120: \quad +2.548 \quad -2.714$$
$$M = 250: \quad +4.702 \quad -4.390$$

9.4.2 Systematic uncertainties overview

Systematic uncertainties for this analysis can be separated in two categories: first category is the "standard" set of systematics affecting the signal yield. Second category are uncertainties related to the background determination procedure.

The main source of systematics of the event yield comes from uncertainties related to jets reconstruction and b-tagging. The second source is the trigger turn-on efficiency, given the rather low thresholds used in the event selection. Other sources include uncertainties on integrated luminosity, PDF

modeling, and lepton-identification. The following uncertainties for the signal event yield have been considered:

- **Trigger systematics**: the trigger turn-on applied on the signal have been derived from data, along with its uncertainties which is mostly coming from the limited statistics of the pre-scaled trigger used as reference sample. We estimate this uncertainties $\approx 3-5\%$;
- **b-tagging efficiency**: this contribution has been studied in detail in a dedicated work [13], using a b-enriched sample from top decay. The scale factor between MC and Data has been included in the efficiency estimated from the MC, and its error is used as a systematics: $\approx 4\%$ per jet, so $\approx 12\%$ for three jets.
- **Jet Energy Scale**: the uncertainty in the Jet Energy Correction (JEC) was estimated by the standard procedure of scaling up and down of the energy of all the jets in each event. Relative change in the amount of the events passing our off-line cuts was $^{+2.5}_{-3.1}\%$
- **Jet Energy Resolution**: to estimate the uncertainty from the Jet Energy Resolution (JER), the momenta of each generated jet were randomly changed according to a corresponding probability distribution for given p_T and η of the jet. The procedure was repeated several times, resulting in different event yields. For the estimate of the JER associated uncertainty we use the standard deviation of the sample: $\pm 1.9\%$
- **Muon momentum scale and resolution**: small $\approx 0.2\%$ and $\approx 0.6\%$ [14];
- **PDF uncertainties**: estimated by reweighting signal MC by the uncertainties of the eigenvectors of covariant matrix of the original PDF. For $M_H = 120$ GeV: $^{+2.5}_{-2.7}\%$; for $M_H = 250$ GeV: $^{+4.7}_{-4.4}$;
- **Integrated luminosity**: 2.2% [15]

Systematic uncertainties related to the background determination procedure arise both from the shape determination and from extraction of absolute normalization.

The background shape is extracted from data using two practically independent methods – one is the matrix method, that uses b-tagging efficiencies

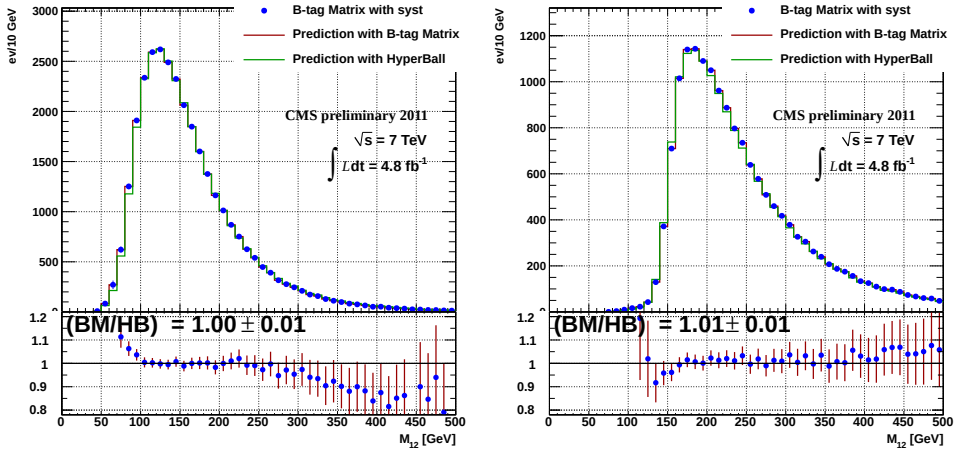


Figure 9.9: Background shape prediction by the b-tagging matrix and "Hyperball" methods.

derived from Monte Carlo and flavour fractions obtained from the template fits to the b-tagging sensitive variables. The hyperball method extract the multidimentional parametrization of probability to find a b-tagged jet in the control sample using a variant of nearest-neighbour algorithm. The shape uncertainty is extracted from comparison of these two background prediction methods – the uncertainty is included in the fitted shape on a bin-by-bin basis.

The normalization uncertainty in turn has two sources: extraction of the normalization factor from control sample and extrapolation of this factor to the signal region. The first uncertainty is shown on Figure 9.7 as the uncertainty in Seen/Pred fraction, which was used as the normalization factor. A systematic error of about $(0.8 \div 0.7\%)$ was associated to that source.

The uncertainty, associated to the extrapolation from control to signal region, is obtained by looking at the QCD Monte Carlo simulation in both regions. The ratio of Seen/Pred ratios:

$$\frac{\text{Seen}_{\text{signal region}}/\text{Pred}_{\text{signal region}}}{\text{Seen}_{\text{control region}}/\text{Pred}_{\text{control region}}}, \tag{9.9}$$

is equal to 1.01 ± 0.042 for the low mass and 1.02 ± 0.05 for high mass region. The bias is accounted for in the limit calculation and the uncertainty is taken as a systematic error of the extrapolation.

9.5 Results and interpretation

In this analysis, the decay products of the $H \to b\bar{b}$ MSSM Higgs boson decay are assumed to give two leading jets of the event. The two jets, therefore, allow us to obtain the kinematics of the decaying Higgs particle. The reconstructed invariant mass of the Higgs candidate is expected to be peaked around the true mass of the resonance, distinguishing it from the broader background contribution.

In order to extract the possible contribution of a MSSM Higgs boson decay in our selected data, we perform a binned likelihood fit (8.24) to the invariant mass distribution of the leading two jets in the events. Events in the fitted data should have at least three b-tagged jets and must satisfy not only the event selection criteria discussed in the Section 9.2, but also should lie in the signal region according to the likelihood discriminants, defined in the Section 9.3.1. Since there are two discriminants for low-mass and high-mass regions, two different data samples was used in the likelihood fits for corresponding MSSM Higgs mass hypothesis.

The two control regions of the data was used for background prediction – the prediction actually used in the fit is produced by the matrix method as described in Section 9.3.1. The prediction of the Hyperball method, described in the Section 9.3.2 was used to estimate the systematic uncertainties, as discussed in the previous section.

Shapes of the signal and predicted background are shown on the Figure. 9.10 for the low-mass region and on Figure. 9.11 for the high-mass region. Signal shapes are shown for different MSSM Higgs m_A mass hypothesis for m^h_{max} benchmark scenario with $\tan\beta = 30$.

9.6 Conclusions

Figure 9.12 on the left shows the derived CL_s limits on the cross-section of the production of neutral supersymmetric Higgs particles in association with two b-quarks, times branching ratio of their decays into another pair of b-quarks. The data were collected during 2011 by the CMS experiment at the LHC, corresponding to a total integrated luminosity of 4.8 fb^{-1}, with the use of a semileptonic trigger.

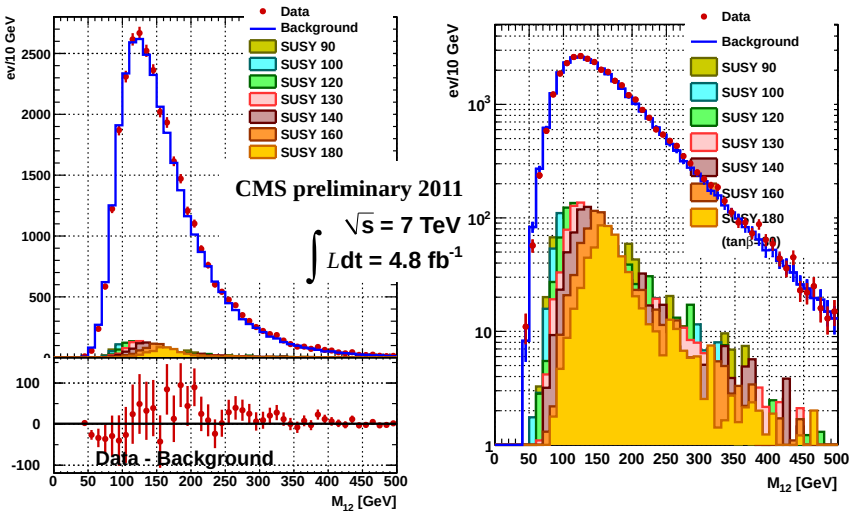


Figure 9.10: Predicted background in the signal region, for low mass range ($M_A < 200$ GeV); the expected signal for different M_A and for $\tan\beta = 30$ is also plotted. Linear scale on the left, logarithmic on the right.

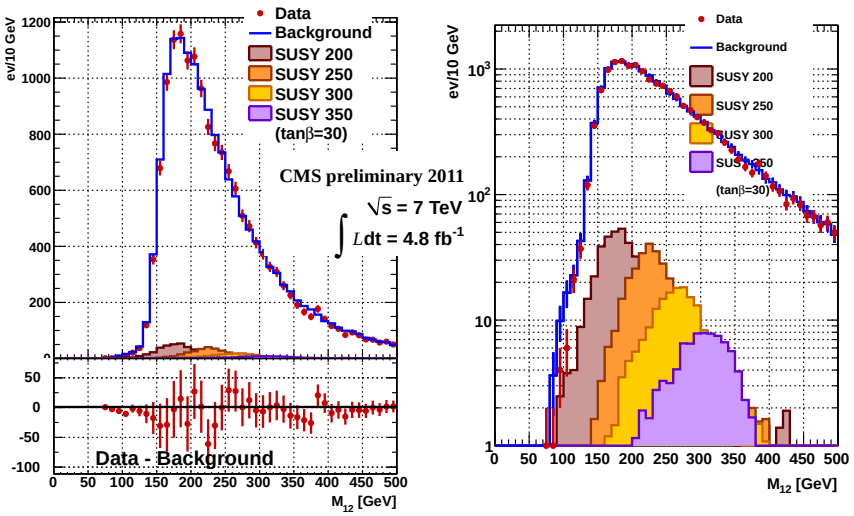


Figure 9.11: Predicted background in the signal region, for high mass range ($M_A > 200$ GeV); the expected signal for different M_A and for $\tan\beta = 30$ is also plotted. Linear scale on the left, logarithmic on the right.

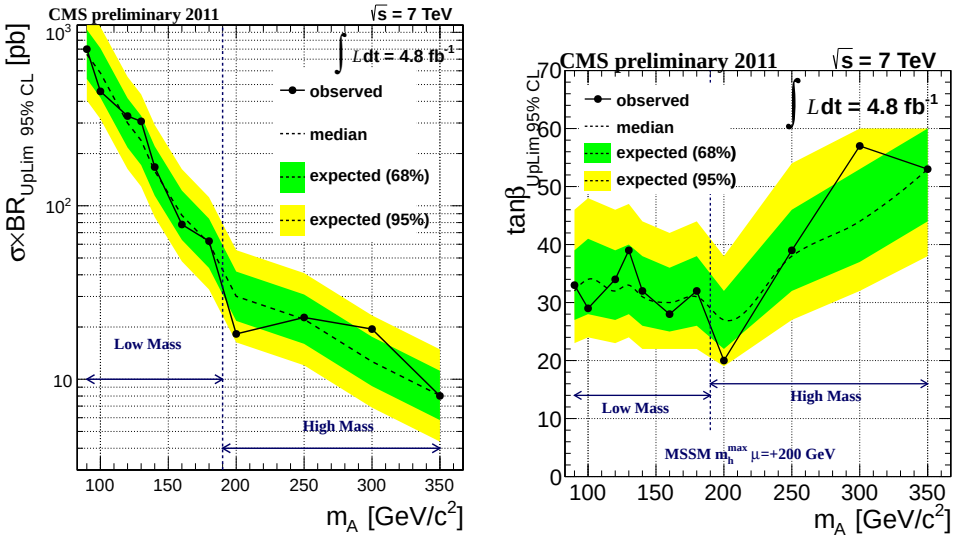


Figure 9.12: CL_s upper limit on the signal cross-section (left) and on corresponding tan β value (right).

The data shows no significant excess with respect to the expected SM multijet QCD background, predicted by means of two independent data driven methods. Since the $\sigma \cdot Br$ scales with the MSSM tan β parameter, the limit can be interpreted as an experimental bound on tan β parameter (shown on the right) in the m_h^{max} scenario.

The results of the study was published as the CMS Physics Analysis Note [16]. The combination with the all-hadronic final state search [**HIG12026**] of similar channel was performed and published in [17]. The combined semileptonic + all-hadronic exclusion in $m_A \otimes \tan\beta$ plane is shown on Figure 9.13 together with previous exclusions from LEP [18] and Tevation [19] searches.

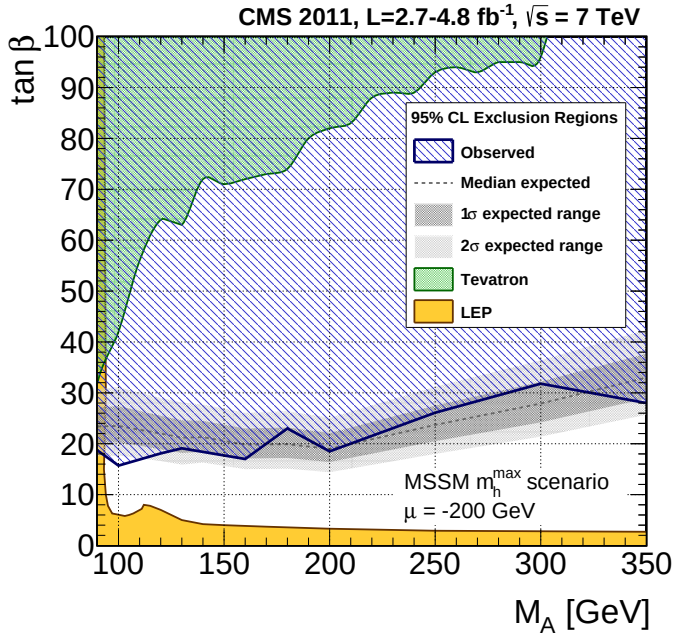


Figure 9.13: Limits on $\tan\beta$ as a function of M_A for the m_h^{max} MSSM scenario. Combination of semileptonic and all-hadronic searches at CMS are shown together with previous exclusions from LEP and Tevatron searches.

Bibliography

[1] Marcela S. Carena et al. "Suggestions for benchmark scenarios for MSSM Higgs boson searches at hadron colliders". *Eur.Phys.J.* C26 (2003), pp. 601–607. arXiv:hep-ph/0202167 [hep-ph].

[2] Marcela S. Carena et al. "MSSM Higgs boson searches at the Tevatron and the LHC: Impact of different benchmark scenarios". *Eur.Phys.J.* C45 (2006), pp. 797–814. arXiv:hep-ph/0511023 [hep-ph].

[3] CMS Collaboration. "Performance of the b-jet identification in CMS" (2011).

[4] S. Amerio, T. Dorigo, and I. Lazzizzera. "Hyperballs User Guide." (2004).

[5] S. Amerio et al. "Narrowing the Higgs mass peak: the hyperball algorithm." (2003).

[6] Roger Barlow. "Systematic errors: Facts and fictions" (2002), pp. 134–144. arXiv:hep-ex/0207026 [hep-ex].

[7] Sergey Alekhin et al. "The PDF4LHC Working Group Interim Report" (2011). arXiv:1101.0536 [hep-ph].

[8] Michiel Botje et al. "The PDF4LHC Working Group Interim Recommendations" (2011). arXiv:1101.0538 [hep-ph].

[9] Pavel M. Nadolsky et al. "Implications of CTEQ global analysis for collider observables". *Phys.Rev.* D78 (2008), p. 013004. arXiv:0802.0007 [hep-ph].

[10] A.D. Martin et al. "Parton distributions for the LHC". *Eur.Phys.J.* C63 (2009), pp. 189–285. arXiv:0901.0002 [hep-ph].

[11] Richard D. Ball et al. "Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology". *Nucl.Phys.* B849 (2011), pp. 296–363. arXiv:1101.1300 [hep-ph].

[12] T. Hahn et al. "FeynHiggs 2.7". *Nucl.Phys.Proc.Suppl.* 205-206 (2010), pp. 152–157. arXiv:1007.0956 [hep-ph].

[13] "Performance of the b-jet identification in CMS" (2011).

[14] Serguei Chatrchyan et al. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV". *JINST* 7 (2012), P10002. arXiv:1206.4071 [physics.ins-det].

[15] CMS Collaboration. "Absolute Calibration of the Luminosity Measurement at CMS: Winter 2012 Update" (2012).

[16] CMS Collaboration. "MSSM Higgs production in association with b-quarks - semi leptonic" (2012).

[17] CMS Collaboration. "Search for a Higgs boson decaying into a b-quark pair and produced in association with b quarks in proton-proton collisions at 7 TeV". *Phys.Lett.* B722 (2013), pp. 207–232. arXiv:1302.2892 [hep-ex].

[18] ALEPH Collaboration. DELPHI Collaboration. L3 Collaboration. OPAL Collaboration. LEP Working Group for Higgs Boson Searches. "Search for neutral MSSM Higgs bosons at LEP". *Eur.Phys.J.* C47 (2006), pp. 547–587. arXiv:hep-ex/0602042 [hep-ex].

[19] CDF Collaboration. D0 Collaboration. "Search for Neutral Higgs Bosons in Events with Multiple Bottom Quarks at the Tevatron". *Phys.Rev.* D86 (2012), p. 091101. arXiv:1207.2757 [hep-ex].

Chapter 10

Search for SM-like Higgs boson in $H \to ZZ \to 2\ell 2q$ channel

Standard Model Higgs boson decaying into a pair of Z bosons is one of the most promising discovery channels at the LHC. A suitable Higgs boson candidate was already discovered with a mass around 126 GeV [1, 2]. The results of this chapter is to be understood in a context of further searches of heavier Higgs-like resonances. Details on the BSM interpretations of the searches will be discussed in the next chapter. Here we will refer to the heavier Higgs-like particle simply as "the Higgs boson".

Depending on the types of subsequent decays of the Z bosons, there are several major classes of studied $H \to ZZ$ channels at the CMS (here ℓ means either e or μ):

$$H \to ZZ \to 4\ell$$
$$H \to ZZ \to 2\ell 2\ell'$$
$$H \to ZZ \to 2\ell 2\tau$$
$$H \to ZZ \to 2\ell 2\nu$$
$$H \to ZZ \to 2\ell 2q$$

Relative sensitivities of these searches strongly depend on the Higgs mass m_H. Generally there are two major regions in the m_H: the so-called "high mass" region with $m_H > 2m_Z$ and the "low mass region" with $m_H < 2m_Z$. In this chapter, the search for SM Higgs boson in $H \to ZZ \to 2\ell 2q$ channel will be discussed. At higher values of m_H, the channel benefits from larger $Z \to q\bar{q}$ branching ratio, but suffers from reduced jet resolution as the Higgs

mass gets lower. In the low mass region one of the Z bosons is off-shell and the search is completely unable do resolve the virtual Z^*.

The search targets at the reconstruction of all four objects, obtaining the full kinematic information about the Higgs decay. As with the analysis in the previous chapter, the main observable of this analysis is the invariant mass of the reconstructed scalar particle. The strategy is to collect all the decay products (two leptons and two jets) and reproduce the kinematics of the decaying Higgs boson. The invariant mass $M_{\ell\ell jj}$ of the $\ell^+\ell^- q\bar{q}$ system is then tested for consistency with Higgs boson signal hypothesis as described in the Section 8.6.

The dominant background processes for this channel are: $Z+$jets production from Drell-Yan process, top-quark pair production ($t\bar{t}$) and diboson production (ZZ, WZ and WW events). The $M_{\ell\ell jj}$ distribution of all the backgrounds is non-resonant, which is useful for signal event isolation. As in the previous chapter the background predictions are, to various degrees, extracted from data: $Z+$jets simulated background shape is reweighed event-by-event to better fit the data, the $t\bar{t}$ distribution is obtained directly from data using the $e^{\pm}\mu^{\mp}$ events.

10.1 Data and simulated samples

Most of the analysis presented in this chapter is preformed with a sample of proton-proton collisions at the center-of-mass energy $\sqrt{s} = 8$ TeV and corresponding to an integrated luminosity of $19.6\,\text{fb}^{-1}$ recorded by the CMS at the LHC during the year 2012. Only data that pass the strict quality requirements imposed by the CMS central certification team are used in the analysis (more technically, the latest available reprocessings and official JSON files are utilized for each data taking period).

Three primary datasets were used in the analysis: `DoubleMu`, `DoubleElectron` and `MuEG` – the last one contains $e^{\pm}\mu^{\mp}$ events for $t\bar{t}$ background determination. The events are collected using a number of un-prescaled and prescaled triggers, like `HLT_Mu17_Mu8` for DoubleMu and `HLT_Ele17_CaloIdT_TrkIdVL_CaloIsoVL_TrkIsoVL_Ele8_CaloIdT_TrkIdVL_CaloIsoVL_TrkIsoVL` for DoubleElectron. The event selection requirements of the analysis are tighter than those of the trigger.

Signal events are generated using POWHEG [3–5]. In signal and background Monte Carlo samples the simulated detector response was produced using the CMS detector model implemented with GEANT4 [6] toolkit. Same reconstruction algorithms was applied both for simulations and for data. The reconstruction is handled by the official CMS software CMSSW, release 5_3_X.

The Z + jets background samples was generated using MADGRAPH V5 [7], imposing a high mass of dilepton pair ($m_{\ell\ell} > 50$ GeV). Although $t\bar{t}$ background is extracted from data, two simulated samples using PYTHIA 6.4.22 [8] and POWHEG [3–5] were produced to cross check the data-driven approach.

Leptonic triggers was used in the data collection, but the trigger requirements was not emulated exactly in the simulations. The simulated events are reweighted according to the event probability to pass the trigger cuts. Corresponding trigger efficiency and MC scale factor tables (as a function of the lepton η and E) for leptons are obtained using the tag-and-probe technique (see Section 6.5.3).

10.2 Physics objects and event reconstruction

10.2.1 Lepton reconstruction

Muons used in the analysis are GlobalMuons satisfying *tight working point* muon ID criteria (see Table 6.2),. Muons are required to be isolated, using the standard tight isolation criteria (6.11). The PileUp corrected isolation (6.10) was used as recommended by the muon POG.

Kinematically, muons should be in the tracker acceptance region $|\eta| < 2.4$ and satisfy extra transverse momenta constraints: $p_T > 40$ GeV for the leading lepton and $p_T > 20$ GeV for the subleading one.

Electrons are the GSF-reconstructed electrons (see Section 6.5.2). Cut-based electron ID was used in this analysis with *loose working point* criteria (see Table. 6.3) – in our case this constraint already has very good selection efficiency and background rejection, so going to tighter electron ID was unnecessary. Further increase in the reconstruction quality is achieved by requiring the ECAL supercluster to lie withing ECAL acceptance $|\eta| < 2.5$, while excluding the barrel-endcap transition region $|\eta| \notin [1.4442, 1.566]$.

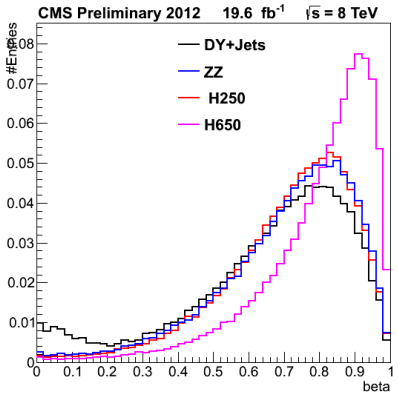


Figure 10.1: Jet β distributions for signal and background events.

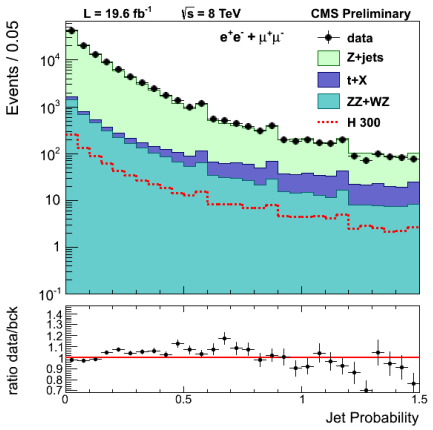


Figure 10.2: Jet Probability tagger distribution.

Similarly to muons, electrons should be isolated (6.12), and have transverse momenta $p_T > 40, 20$ GeV for leading and subleading electron correspondingly.

10.2.2 Jets and b-tagging

Jets used in the analysis are standard Particle Flow jets, reconstructed by anti-k_t algorithm with $\Delta R = 0.5$. All the Jet Energy Corrections are applied (see Section 6.4.3) both for data and simulation. In order to achieve highest possible reconstruction efficiency, jets are required to have $p_T > 30$ GeV and lie within tracker acceptance $|\eta| < 2.4$. As it was discussed in Section 6.4.2, in this region there are two variables β and β^*, (6.3) that are most discriminating against PileUp jets.

It turns out that very good rejection of PileUp jets can be already accomplished by imposing a simple constraint on β variable as demonstrated on Figure 10.1. PileUp jets are removed using a simple $\beta > 0.2$ cut. Such approach, approved by the b-tagging POG, is efficient enough for our purposes and has an advantage of being quite simple as compared to more involved MVA-based PileUp-jet rejection approaches.

The hadronic branchings of the Z boson is almost equally distributed among the u, d, s, c, b quarks, while the most important $Z +$ jets background contains leptonically decaying Z and two associated jets with high transverse

momenta. Production of such jets is expected to be dominated by valence partons of the proton: gluons and u and d quarks. As a result, b-tagging can help us increase the analysis sensitivity in the case of the $H \to ZZ \to 2\ell 2q$ analysis by substantially reducing the $Z + jets$ background.

The b-tagging algorithm (see Chapter 7) used in this analysis is the Jet Probability (JP) tagger (7.1). The reason for this choice (disfavoring more sophisticated algorithms like CSV) lies in the fact that for smaller values of m_H jets are mostly forward. In such "semi-forward region" CSV has certain issues: the c-jets fake rates are unacceptably large – that would considerably reduce the analysis sensitivity, especially affecting the 1 b-tag category.

The b-tagging POG provides three working points for JP tagger: loose ($JPL: \ JP < 0.275$), medium ($JPM: \ JP < 0.545$) and high ($JPH: \ JP < 0.79$). Events are classified in three mutually exclusive categories according to the number of b-tagged jets:

- **2 b-tag** category requires at least two b-tagged jets – one satisfying JPM working point and at least one another passing JPL constraint.
- **1 b-tag** category contains jets with exactly one b-tagged jets with JPL
- **0 b-tag** category containing the rest of the events.

Figure 10.2 shows the distribution for JP discriminant after all other selection cuts both for data and simulation. The simulation distribution was rescaled using the Moriond13 prescription for scaling the b-tagging efficiencies: $\langle \epsilon_f \rangle = \langle \epsilon_f^{MC} \rangle \cdot \langle SF_f \rangle$ for flavours $f = b, c, light$. The efficiencies and mistag rates are further improved by accounting for p_T and $|\eta|$ dependence by taking mistag rates $\epsilon_f^{2011}(p_T, |\eta|)$ provided by the b-tag POG for 2011 data and rescaling them to match 2012 averaged efficiencies.

10.2.3 Missing Transverse Energy

An important background for the 2 b-tag category arises from the $t\bar{t}$ production with subsequent $t \to bW \to b\ell\nu_\ell$ decays. The background could, however, be distinguished by a visible Missing Transverse Energy (see Section 6.4.4), carried out by the neutrinos from the W decays. Signal events do not have any physical MET, therefore requiring selection of events with vanishing MET should reduce the $t\bar{t}$ background.

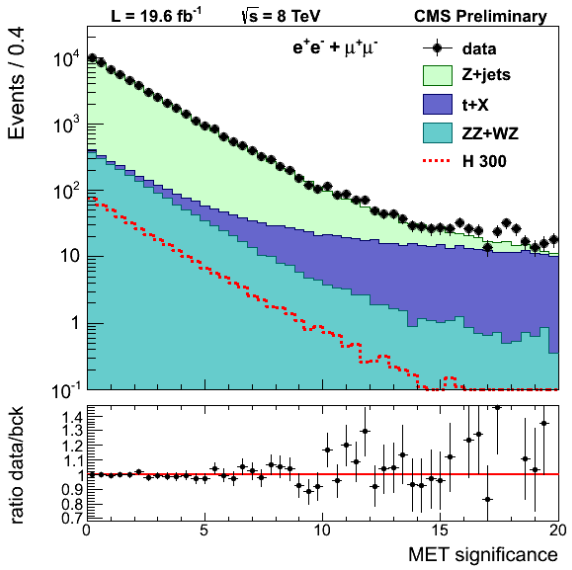


Figure 10.3: Distribution of the λ_{MET} in the selected event. The simulation is scaled by a factor of 1.15 to match the data.

We are interested in events that do not have any *physical* MET, while the observed MET will always be non-zero due to resolution effects. Precise simulation of such effects is very difficult, especially in the presence of PileUp. The MET significance variable λ_{MET} (see eq. (6.9)) is designed to account for the resolution effects, measuring the likelihood to have $\not{E}_T^{true} = 0$ given the observation. It also have another advantage of being much less sensitive to PileUp contamination, since the PileUp contributions turn out to cancel in the likelihood ratio (6.9).

In order to reduce the $t\bar{t}$ background the $\lambda_{MET} < 10$ cut was imposed for all b-tagging categories. Figure 10.2.3 shows the distribution in λ_{MET} both in data and in the simulation. The shape of the distribution in data is well reproduced by the simulation, but an extra global scaling factor of ~ 1.15 was necessary due to possible imperfections in modeling of the jet resolution. The effect of this rescaling on the signal efficiency of the $\lambda_{MET} < 10$ cut is about 0.5%. This variation is included in systematic uncertainties of the analysis.

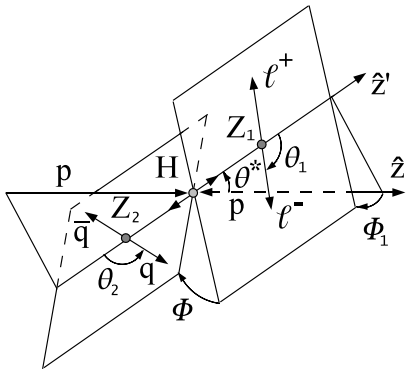


Figure 10.4: Five angular variables describing the $H \to ZZ \to 2\ell 2q$ decay topology.

10.2.4 Reconstruction of Z and Higgs candidates

All the reconstructed leptons and jets are combined in *Z-candidates*. Leptonic Z-candidates are constructed by picking opposite-charged lepton pairs that satisfy $m_{\ell^+\ell^-} \in [76, 106]$ GeV. Hadronic $Z \to q\bar{q}$ candidates are constructed from all jet pairs passing $m_{jj} \in [60, 130]$ GeV. The m_{jj} range is then divided into a signal region, covering $m_{jj} \in [71, 111]$ GeV and the sideband region made of the rest of the events.

Finally, Higgs candidates are assembled from pairs of leptonic and hadronic Z candidates. Sideband $Z \to q\bar{q}$ candidates are used only if there are no hadronic Z candidates in the signal region. In some cases (about 3% af all the events) several Higgs candidates can be built in a single event. The candidate in a highest b-tag category is then selected. If there are several candidates in that category – then the candidate with minimal value of $|m_{\ell\ell} - m_Z| + |m_{jj} - m_Z|$ is chosen.

10.2.5 Angular Likelihood Discriminant

The kinematics of the $H \to ZZ \to \ell^+\ell^- q\bar{q}$ decay chain exhibits some angular correlations that could be used to increase the discriminating power of the analysis. For example the ΔR distance between the products of $Z \to q\bar{q}$ and $Z \to \ell^+\ell^-$ decays should decrease for increasing values of m_H, since the transverse momenta of the bosons increase for heavy Higgs boson, resulting

in boosted kinematics of Z decays. Moreover, there are spin correlations in $H \to ZZ$ and subsequent Z decays that are transformed in the extra angular correlations.

Jets of the Z + jets background do not have any such correlations and this fact can be used to further reduce the Drell-Yan background. It was shown in [9], that there are five angles, that fully describe the $H \to ZZ \to \ell^+\ell^- q\bar{q}$ decay kinematics. These angles are shown on Figure 10.4, all of them are considered in the rest frame of the decaying Higgs boson:

θ^* **angle** is the angle between the proton collision axis and the $H \to ZZ$ decay axis.

θ_1 **and** θ_2 **angles** are the angles between the $H \to ZZ$ decay axis and the axes of the $Z \to \ell^+\ell^-$ and $Z \to q\bar{q}$ decays correspondingly.

Φ_1 **angle** is the angle between $H \to ZZ$ decay plane and the plane of the $Z \to \ell^+\ell^-$ decay.

Φ **angle** is the angle between the two $Z \to \ell^+\ell^-$ and $Z \to q\bar{q}$ planes.

The angular variables $(\theta^*, \theta_1, \theta_2, \Phi_1, \Phi)$ are orthogonal to the invariant masses and transverse momenta of the intermediate resonances and independent on the production mechanism. That property allows us to perform an extra selection based on the event topology in the Higgs boson rest frame, while keeping the distributions in m_{jj}, $m_{\ell\ell}$ and $m_{jj\ell\ell}$ unaffected.

There is no single angular variable with dominating discriminating power – so, to fully exploit information about angular correlations, the likelihood discriminant was constructed using all the five variables presented above.

$$LD(\theta^*, \theta_1, \theta_2, \Phi_1, \Phi; m_{jj\ell\ell}) = \frac{P(signal)}{P(background) + P(signal)} \tag{10.1}$$

The $P(signal)$ distribution is parametrized as a product of an ideal theoretical probability function $P(ideal)$, derived in [9] and acceptance functions $G(x; m_{jj\ell\ell})$ for individual variables:

$$P(signal) = P(ideal) \prod_{\theta^*, \theta_1, \theta_2, \Phi_1, \Phi} G_x(x; m_{jj\ell\ell}) \tag{10.2}$$

$P(background)$ is obtained from the simulation as a product of five one-dimensional distributions (thus neglecting correlations. The $P(background)$ parameterization also depends on $m_{jj\ell\ell}$.

All selected events are required to have a value of the likelihood discriminant larger than $LD > 0.5$, which reduces the $Z +$ jets background by a factor of two, while retaining a signal efficiency of at least 80% for all m_H.

It is important to stress that the LD selection relies on the spin and parity properties of the produced resonance. The angular selection is optimized to observe the CP-even spin-0 SM-like Higgs boson.

10.3 Signal model

For higher SM Higgs masses, the width of the resonance becomes very large compared to the mass. Quality of the so-called Zero Width Approximation degrades as $\sim O(\Gamma_H / M_H)$ and becomes unsatisfactory, requiring better description of the Higgs lineshape (distribution in invariant mass). Better description of the lineshape, known as *Complex Pole Scheme* (CPS) was proposed in [10]. The CPS total cross-section was computed by the Higgs Cross-Section Working group [11]. The `POWHEG` simulated signal samples are reweighted so that signal shape matches the CPS shape.

Another issue is the interference between $gg \to H \to ZZ$ signal and $gg \to ZZ$ non-resonant background (like t-box diagrams). Interference starts to contribute significantly to the Higgs lineshape at higher m_H, as discussed in [12]. Unitarity requires the interference to be destructive at higher values of the m_{ZZ} above the mass peak. Below the mass peak the interference is constructive, so the correction to a total cross-section due to interference is quite small – it is of order of $1 \div 2\%$. But it visibly affects the Higgs lineshape, which is especially important for studies, that are sensitive to the invariant mass distribution.

The situation is complicated by the fact, that the results of the interference calculations are available only at the leading order (LO). A certain procedure was developed [12] for combination of the LO interference result with the Higgs production cross-section obtained at NNLO and for estimation of the associated uncertainty. The details of this procedure, extended to handle BSM Higgs scenario will be presented further in Chapter 11.

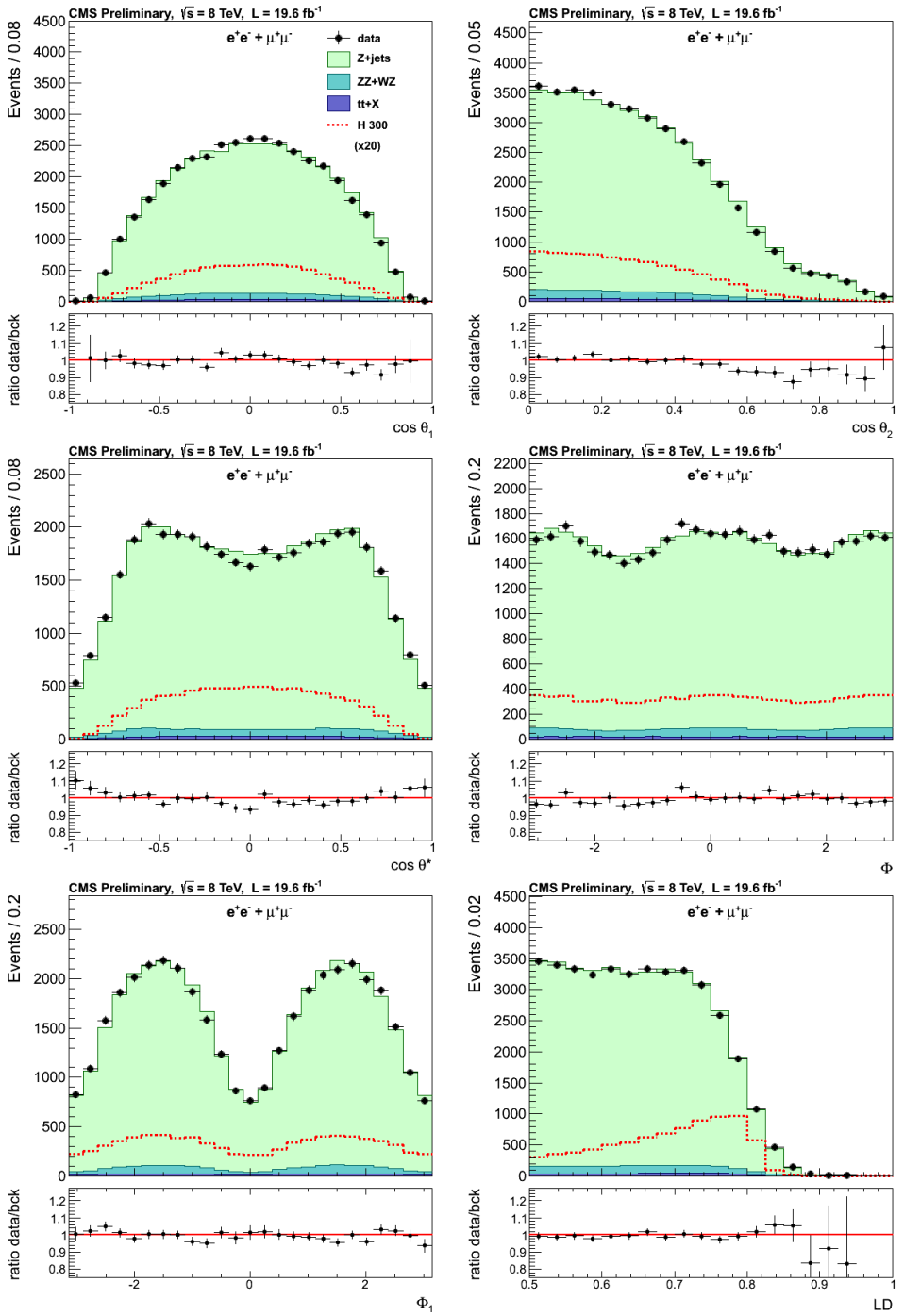


Figure 10.5: Distributions of the angular variables and the likelihood discriminant after the final selection. Data and MC are shown. Background distributions are normalized from data using sidebands.

10.4 Background determination

After the full event selection, there are three processes, that can be considered as a significant background sources: double boson production (ZZ and WZ mainly), $Z+$jets and production $t\bar{t}$. For the diboson production we rely on the MC simulation, while the two other backgrounds are either normalized to data in sidebands ($Z+$jets) or extracted directly from data using the sidebands ($t\bar{t}$ production).

$Z+$jets Monte Carlo samples are produced with Madgraph [13]. Four samples of $Z+n$jets events for $n=1,2,3,4$ was produced. We've observed some differences between data and simulations, but the mis-modeling of the p_T spectrum of the $\ell\ell jj$ system is the same in the signal region and in the sidebands. Using data and Monte Carlo p_T distributions in the sidebands, we reweighted the simulated $Z+$jets distribution in the signal region. The contamination of $t\bar{t}$ and diboson events was first subtracted from sideband data sample.

Determination of the $t\bar{t}$ background relies on the fact that the top-quark leptonic decays are almost completely lepton flavour symmetric, so there is no preference between $e^{\pm}\mu^{\mp}$ and $\mu^+\mu^- + e^+e^-$ events in the $t\bar{t}$ processes. Notice that this argument extends to other small backgrounds (like $WW+$ jets, $Z \rightarrow \tau^+\tau^- +$jets, single top, fakes) that do not distinguish between lepton flavors.

Studies of $t\bar{t} \rightarrow 2\ell 2\nu + X$ Monte Carlo samples show that the lepton flavor symmetry works very well for the shape. Normalization of the distribution agrees within Monte Carlo statistical errors.

10.5 Systematics

Systematic uncertainties for this analysis can again be subdivided in the uncertainties on signal yield and uncertainties of the data-driven background determination.

- **Jet Energy Scale and Resolution**: the uncertainty of the jet reconstruction mainly comes from the Jet Energy Scale factors. The JES variation in $\pm 1\sigma$ range changes the signal yeld differently for different mass hy-

	PDF			
m_H	CTEQ66	MSTW2008NLO	NNPDF2.1	Total
200	+0.6%/-0.7%	-0.2%/-0.5%	+0.8%/+0.2%	+0.8%/-0.7%
400	+0.8%/-1.0%	+0.6%/+0.2%	+1.4%/+0.75%	+1.4%/-0.8%
600	+0.8%/-1.1%	+0.8%/+0.4%	+1.5%/+0.9%	+1.5%/-1.1%
800	+1.5%/-2.0%	+1.5%/+0.7%	+2.7%/+1.4%	+2.7%/-2.0%
1000	+2.6%/-3.2%	+2.5%/+1.2%	+4.3%/-2.4%	+4.3%/-3.2%

Table 10.1: Systematic effects due to PDF uncertainties.

pothesis:

m_H	$+1\sigma$	-1σ
250	4.3%	-4.2%
300	1.3%	-1.3%
400	1.2%	-1.2%
600	-0.8%	0.9%

- **PileUp reweighting**: as it was discussed in the Chapter 6.2, it is a standard practice to reweight the simulated samples to that the number of interactions per bunch crossing matches the distribution in data. The systematic error, arising from such reweighting was estimated by changing the number of true interactions following the recommendations in [14]. The consequent change in the signal efficiency is less than 1% for 0 and 1 b-tag categories and about 2% for 2 b-tag category.

- **B-tagging efficiency**: The scale factors between data and Monte Carlo, mentioned in Section 10.2.2, are correcting for mismatch in tagging efficiencies and fake rates in data and simulation. The associated systematic effect was estimated by varying the efficiencies in their corresponding uncertainty ranges.

- **MET uncertainty**: the λ_{MET} scaling factor due to mis-modeling discussed in Section 10.2.3 affects the signal efficiency by about 0.5%. Rest of the possible uncertainty comes from other sources, like jet reconstruction and PileUp.

- **Lepton trigger, ID and isolation**: computed using tag-and-probe technique. The total uncertainty for muons is 2.7% and 2% for electrons.

- **Lepton scale corrections**: very small, below 1%
- **Integrated luminosity**: 4.4% [15]
- **PDF uncertainties**: estimated by following PDF4LHC [16] recommended procedure. The uncertainties do depend on the Higgs mass hypothesis as shown on the Table 10.1.
- **Lineshape uncertainty**: the systematic effect due to lineshape approximation discussed in the Section 10.3 was estimated by varying the lineshape prediction with corresponding K-factor reweighting. The effect is negligible below $m_H = 400$ GeV and rises to 3% at 600 GeV.

10.5.1 Background determination systematics.

Many of the systematic uncertainties above are propagating to the uncertainty on background determination, affecting the shape and normalization of the background $M_{\ell\ell jj}$ distribution.

Lepton trigger/reconstruction and jet energy scale uncertainties contribute 2% and 5.5% to normalization uncertainty. The *b*-tagging efficiency scale factor uncertainties has an effect on the normalization of 0.4%, 0.8% and 4.5% for the 0-, 1- and 2-btag categories respectively. The uncertainty on the b-tagging mistag rates introduces an uncertainty in the normalization of 1.9%, 7.8% and 6.2% for the 0-, 1- and 2-btag categories respectively.

The shape of $M_{\ell\ell jj}$ distribution is affected by jet energy scale and $Z+$jets p_T correction uncertainties. The jet energy scale effect varies from 0% at low masses up to 4% at 600 GeV. The p_T reweighting uncertainty is evaluated by comparing $M_{\ell\ell jj}$ distribution with and without the correction. The effect goes up to 3% at high masses.

Some residual differences in the 4-body invariant mass distribution between data and prediction in the sideband region are observed. The difference is taken as an additional mass-dependent systematic uncertainty.

10.6 Results and interpretation

Like in the analysis of the Chapter 9, the fully reconstructed Higgs boson kinematics allows us to use the invariant mass distribution of the reconstructed Higgs candidate in a binned likelihood fit. The signal distribution is

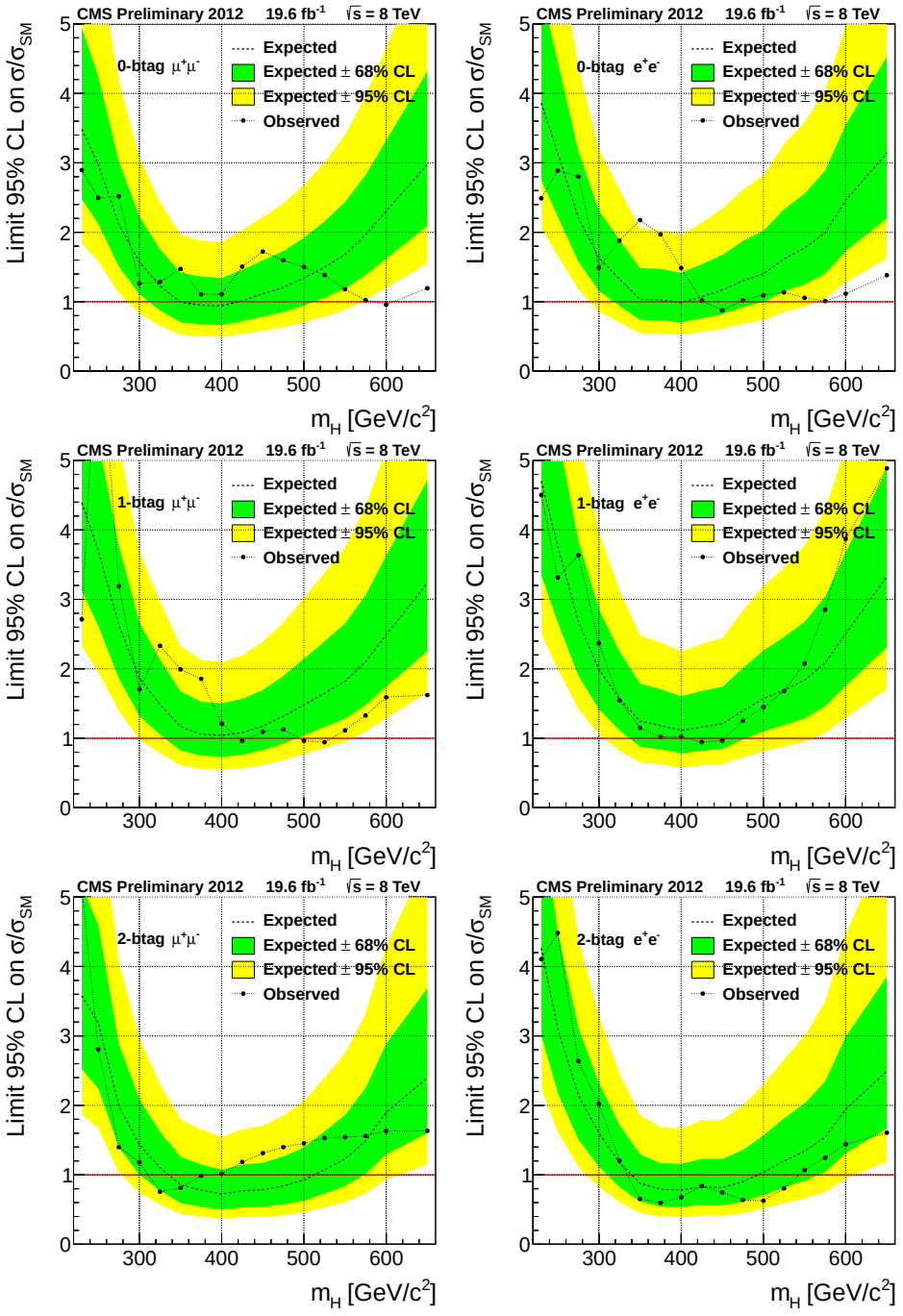


Figure 10.6: Expected and observed limits for six individual channels: muons on the right, electrons on the left. Each row corresponds to the 0, 1 and 2-btag even categories.

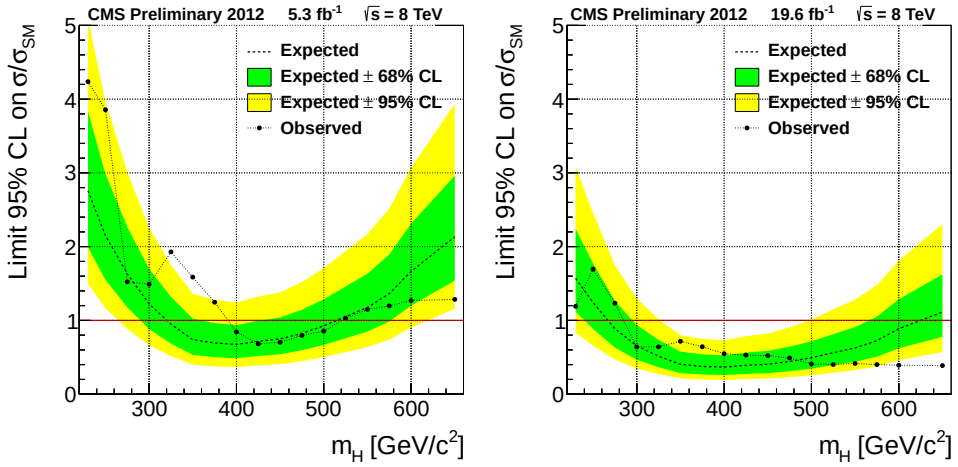


Figure 10.7: Expected and observed limits for integrated luminosities of 5.3 fb^{-1} (left) and 19.6 fb^{-1} (right) for $\sqrt{s} = 8$ TeV collisions.

expected to be peaked around the true m_H, while the background has a much broader distribution.

The calculation of the corresponding limits was performed independently for six distinct classes of events: electrons and muons in 0, 1, and 2 b-tag categories. For each of the six classes, the expected upper limits on the SM Higgs boson production cross section are calculated as a function of the m_H hypothesis. Details of the statistical procedure are described in the Chapter 8.6. It is implemented as an official tool, provided by the CMS Higgs combination group [17].

10.6.1 Interpolation and combination of the results

The observed limit on signal strength μ is determined for m_H hypotheses between 230 GeV and 650 GeV on a quite sparse grid, motivated by the low resolution of the analysis.

Figure 10.7 shows the expected and observed limits, combining all six channels for all the considered mass hypotheses for the 5.3 fb^{-1} ICHEP data (left), and for the full dataset (right) recorded during 2012 at 8 TeV, corresponding to a luminosity of 19.6 fb^{-1}.

Other CMS searches have produced limits on a much finer grid of resonance m_H hypotheses. Also the study of the $H \to ZZ \to 2\ell 2q$ channel

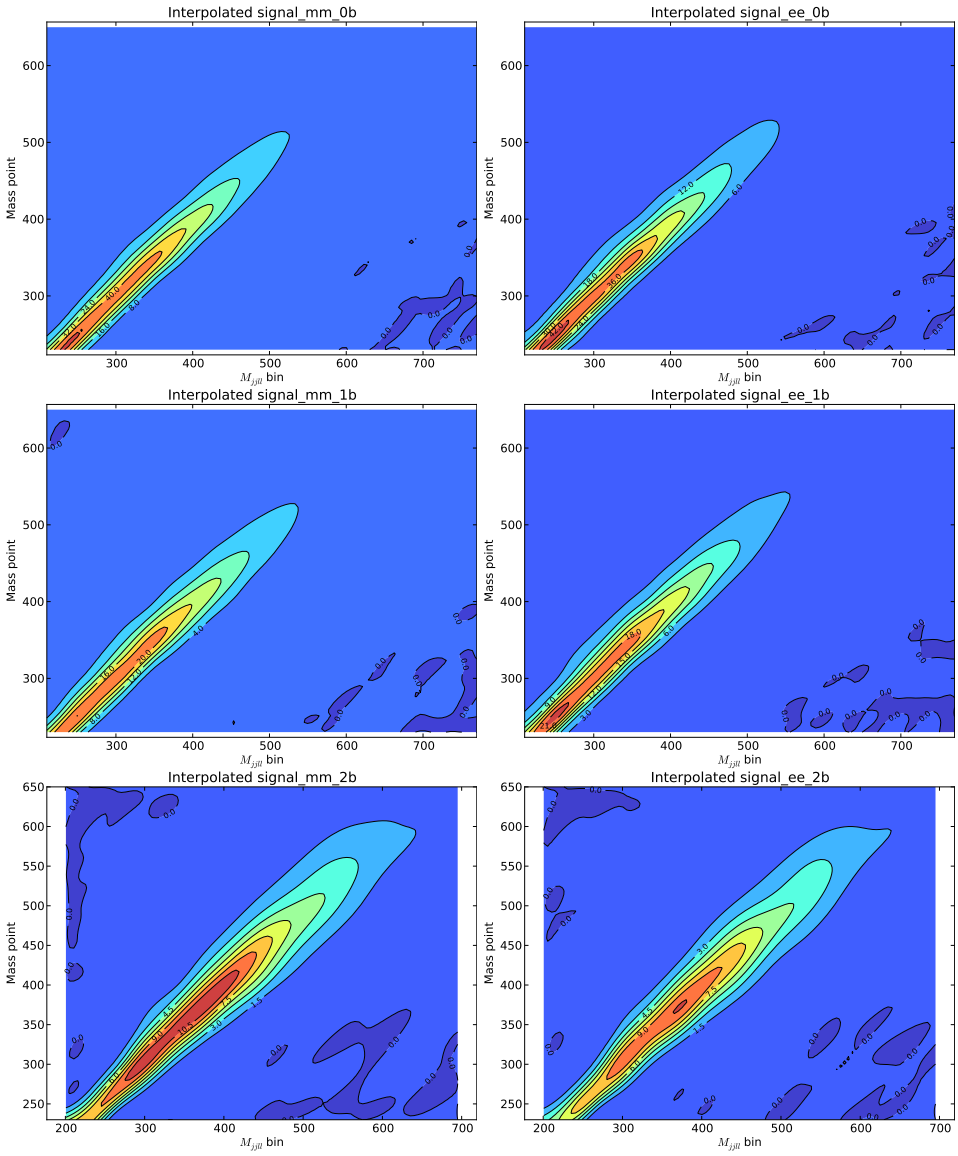


Figure 10.8: Signal histograms, interpolated as a function of m_H hypothesis for the six categories: $(ee + \mu\mu) \times$ (0-b-tag, 1-b-tag, 2-b-tag).

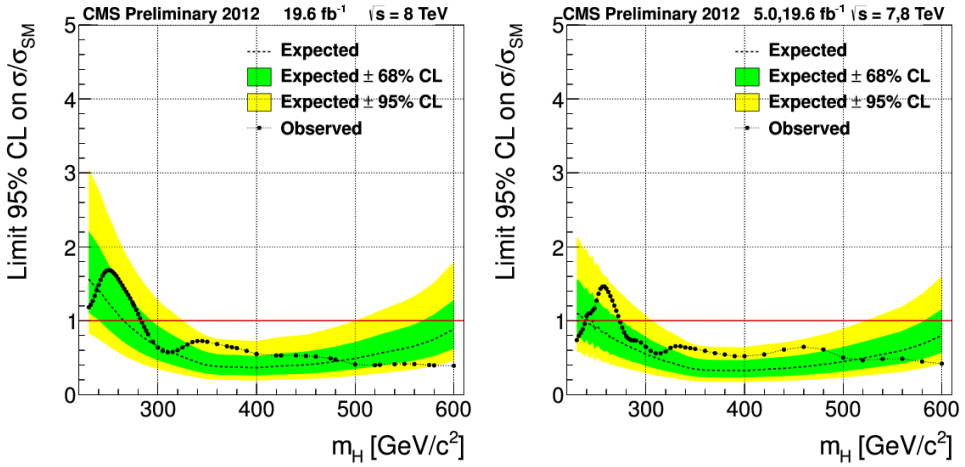


Figure 10.9: Interpolated limit for full $\sqrt{s} = 8$ TeV data taking (left) and for $\sqrt{s} = 7, 8$ TeV combination (right).

at $\sqrt{s} = 7$ TeV was performed with a different m_H grid with many non-overlapping points. In order to combine our results with these searches, we interpolate the histograms used for the signal hypothesis to those points where no simulation is available.

For the interpolation we've used the Radial Basis Function [18] spline. The spline was used to fit the expected signal shapes on a $m_H \otimes m_{\ell\ell qq}$ plane individually for each of the six categories. The interpolated splines are demonstrated on Figure 10.6.1. The fitting was cross-checked by removing a histogram from the spline and looking on the way it is reproduced by the interpolation.

10.7 Conclusions

Figure 10.9 shows the combined and interpolated result of the search of a SM-like Higgs boson, decaying into a pair of Z-bosons, which subsequently decay into a pair of leptons and into a pair of quarks, $H \to ZZ \to 2\ell 2q$. Figure on the left shows the limit derived from 2012 data corresponding to an integrated luminosity of $19.6\, fb^{-1}$ at proton-proton collisions at center-of-mass energy of $\sqrt{s} = 8$ TeV. Figure on the right shows a combination of the present and previous search in 2011 data that corresponds to $5.0\, fb^{-1}$ of in-

tegrated luminosity. The combination of the two results allows us to exclude the existence of a SM-like Higgs resonance in the mass range between 275 and 600 GeV.

The preliminary results of the study (5.3 fb^{-1} at $\sqrt{s} = 8$ TeV) was published in [19]. The combined paper documenting joined effort of all the $H \to ZZ$ searches at high masses is in preparation. Some extra developments in the $H \to ZZ \to 2\ell 2q$ analysis are planned to be finished. Due publication the following items are planned to be completed:

- Two Higgs production channels significantly contribute to the signal rate: the gluon fusion production and the Vector Boson Fusion (VBF). The gluon fusion is dominates the production cross-section, but VBF still has a visible contribution.

  The Vector Boson Fusion production can and should be treated separately, increasing the sensitivity by exploiting the information from the associated jets.

- The main analysis is performed in the m_H range from 220 GeV to 650 GeV – the lower bound is determined by the two on-shell Z's threshold, while masses above 650 GeV jets get highly boosted and merge in a single "fat jet"; dedicated studies of the merged jets are required in that case.

- Reinterpretations of the search results in terms of some Beyond Standard Model extensions are possible without redoing the whole analysis from scratch. Some developments in that direction are reported in the next Chapter.

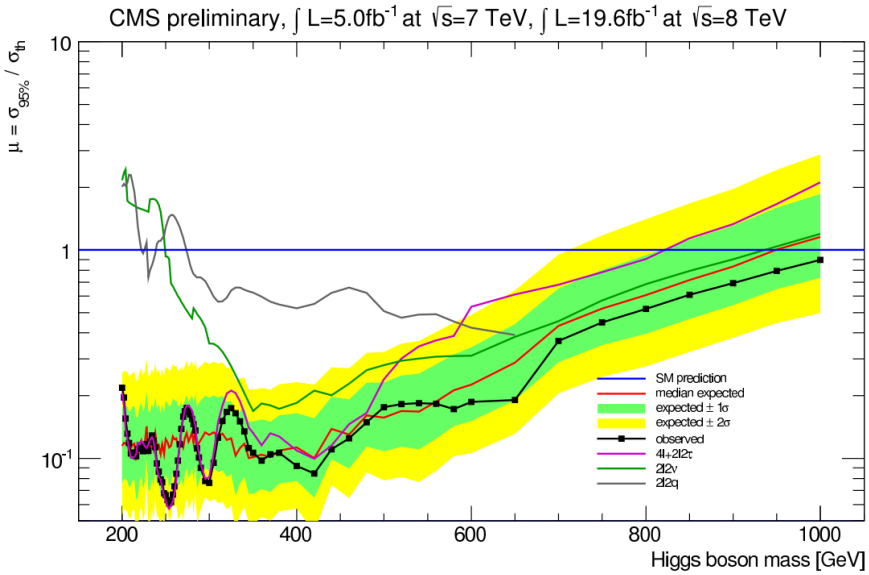


Figure 10.10: Combination of the observed limits of the searches of heavy SM-like Higgs boson decaying in a pair of Z bosons in 4-lepton (both 4ℓ and $2\ell 2\tau$), $2\ell 2\nu$ and $2\ell 2q$ channels. Both $\sqrt{s} = 7$ and 8 TeV data taking periods are combined.

Bibliography

[1] CMS Collaboration. "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$ = 7 and 8 TeV". *JHEP* 1306 (2013), p. 081. arXiv:1303.4571 [hep-ex].

[2] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Phys.Lett.* B716 (2012), pp. 1–29. arXiv:1207.7214 [hep-ex].

[3] Paolo Nason. "A New method for combining NLO QCD with shower Monte Carlo algorithms". *JHEP* 0411 (2004), p. 040. arXiv:hep-ph/0409146 [hep-ph].

[4] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". *JHEP* 0711 (2007), p. 070. arXiv:0709.2092 [hep-ph].

[5] Simone Alioli et al. "NLO vector-boson production matched with shower in POWHEG". *JHEP* 0807 (2008), p. 060. arXiv:0805.4802 [hep-ph].

[6] S. Agostinelli et al. "GEANT4: A Simulation toolkit". *Nucl.Instrum.Meth.* A506 (2003), pp. 250–303.

[7] Johan Alwall et al. "MadGraph 5 : Going Beyond". *JHEP* 1106 (2011), p. 128. arXiv:1106.0522 [hep-ph].

[8] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. "PYTHIA 6.4 Physics and Manual". *JHEP* 0605 (2006), p. 026. arXiv:hep-ph/0603175 [hep-ph].

[9] Yanyan Gao et al. "Spin determination of single-produced resonances at hadron colliders". *Phys.Rev.* D81 (2010), p. 075022. arXiv:1001.3396 [hep-ph].

[10] Stefano Goria, Giampiero Passarino, and Dario Rosco. "The Higgs Boson Lineshape". *Nucl.Phys.* B864 (2012), pp. 530–579. arXiv:1112.5517 [hep-ph].

[11] LHC Higgs Cross Section Working Group et al. "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables". *CERN-2011-002* (CERN, Geneva, 2011). arXiv:1101.0593 [hep-ph].

[12] Giampiero Passarino. "Higgs Interference Effects in $gg \to ZZ$ and their Uncertainty". *JHEP* 1208 (2012), p. 146. arXiv:1206.3824 [hep-ph].

[13] Johan Alwall et al. "MadGraph/MadEvent v4: The New Web Generation". *JHEP* 0709 (2007), p. 028. arXiv:0706.2334 [hep-ph].

[14] CMS Collaboration. "Estimating systematic errors due to pileup modeling". *https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupSystematicErrors* (2012).

[15] CMS Collaboration. "CMS Luminosity Based on Pixel Cluster Counting - Summer 2012 Update" (2012).

[16] Sergey Alekhin et al. "The PDF4LHC Working Group Interim Report" (2011). arXiv:1101.0536 [hep-ph].

[17] CMS Collaboration. "Estimating systematic errors due to pileup modeling". *https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideHiggsAnalysisCombinedLimit* (2012).

[18] "Numerical Recipes 3rd Edition: The Art of Scientific Computing" (Sept. 2007).

[19] CMS Collaboration. "Search for a standard-model-like Higgs boson with a mass in the range 145 to 1000 GeV at the LHC". *Eur.Phys.J.* C73 (2013), p. 2469. arXiv:1304.0213 [hep-ex].

Chapter 11

BSM interpretations of the $H \to ZZ \to 2\ell 2q$ searches

The search, discussed in the previous chapter, is much more sensitive to the Higgs masses above ZZ threshold $m_H > 2m_Z \sim 200$ GeV. The Higgs boson candidate was, however, observed [1, 2] with a mass around 126 GeV. Still, it would be unfair to say that analyses that weren't very sensitive to lower mass ranges didn't play a role in the discovery. The discovery of the Higgs boson at $m_H = 125$ GeV was made in a wider context of the whole possible mass ranges up to $\sim 1TeV$. After all, the combination of the exclusion analyses, showing that there are no extra resonances at other values of m_H, was one of the reasons that allowed us to affirm that this new boson is actually the Higgs boson.

It is important to stress, however, that all these higher-mass searches in every channel was targeting the Standard Model Higgs boson. After the discovery, all that analyses can be considered as searches of a different "SM-like scalar resonance" – boson that would have all the properties (like cross-sections and branching ratios) of the Higgs boson in the Standard Model with different values of m_H. Bearing that in mind, it sounds useful to interpret the already obtained results in terms of more theoretically justified BSM models.

In general, experimental studies of models that extend the Standard Model require redoing of the whole analysis workflow – production of dedicated Monte Carlo samples, optimization of the selection procedure, e.t.c. Some credible way of using all the existing developments to reinterpret the data in

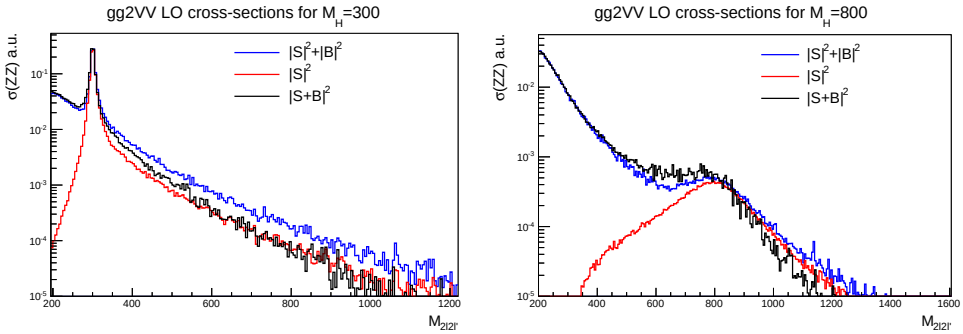


Figure 11.1: $|S|^2_{LO}$,$|S|^2_{LO}$+$|B|^2$, and $|S+B|^2_{LO}$ lineshapes obtained from `gg2VV` package for $m_H = 300$ GeV and $m_H = 800$ GeV Higgs mass hypotheses.

a context of some BSM models could in principle provide exclusions for them with much less effort. BSM models "eligible" for that kind of study should provide a single scalar particle with the same (or quite similar) production kinematics, same CP parity and spin.

In the following we will discuss the current developments on the reinterpretation of the heavy Higgs searches in the $H \to ZZ \to 2\ell 2q$ channel in terms of two of the BSM models: the EWK singlet model and the Two Higgs Doublet Model.

11.1 Electroweak singlet model

Electroweak (EWK) singlet model introduces an extra scalar field ϕ_H, which mixes with the original Higgs scalar, connecting the Standard Model fields to the so-called "hidden" sector. The more detailed description of this model was done in Section 3.2.

For the current discussion the most important model's feature is the relations (3.11). These relations make the reinterpretation of the results of the heavy Higgs searches very straightforward. In this model one just needs to rescale cross-sections and couplings to get the exclusions on the model.

11.1.1 Signal+background interference effects.

Following the procedure in [3], we obtain signal shapes $|S_{LO}|^2$, background shape $|B|^2$, and combined shapes $|S+B|^2_{LO}$. The three kinds of distributions

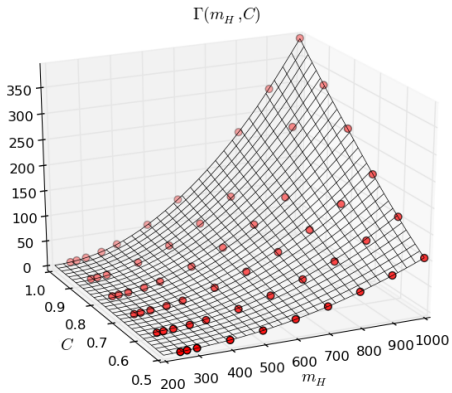


Figure 11.2: Fitted and interpolated values of Γ as a function of m_H and C.

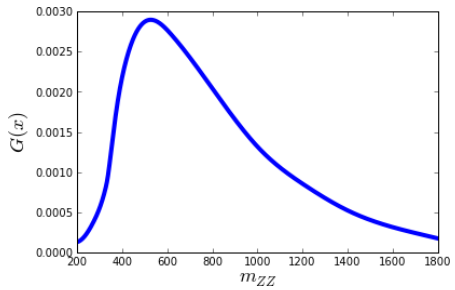


Figure 11.3: The spline function $G(m_{ZZ})$ used in signal fitting function (11.5).

should be calculated for for various Higgs mass hypothesis m_H and anomalous BSM couplings C. The interference contribution I_{LO} is then extracted as:

$$I_{LO} = |S+B|^2_{LO} - |S|^2_{LO} - |B|^2. \tag{11.1}$$

For the calculations we've used `gg2VV` [4, 5], which evaluate leading order cross-sections and event generation for $gg \to H \to ZZ$ and $gg \to H \to WW$ processes. The package allows the user to "switch off" contributions from signal or background and to modify the Higgs mass, width and BSM couplings. Figure 11.1 shows the effect of signal-background interference by overlaying $|S|^2_{LO}+|B|^2$ and $|S+B|^2_{LO}$ lineshapes obtained from `gg2VV` for $m_H = 300$ GeV and $m_H = 800$ GeV.

11.1.2 Lineshape interpolation for BSM searches

As one can see, the results of the interference calculation are available only at the leading order (LO). The way to estimate the next-to-next-to-leading order shape (NNLO) and cross-section was suggested in [3]. Following the

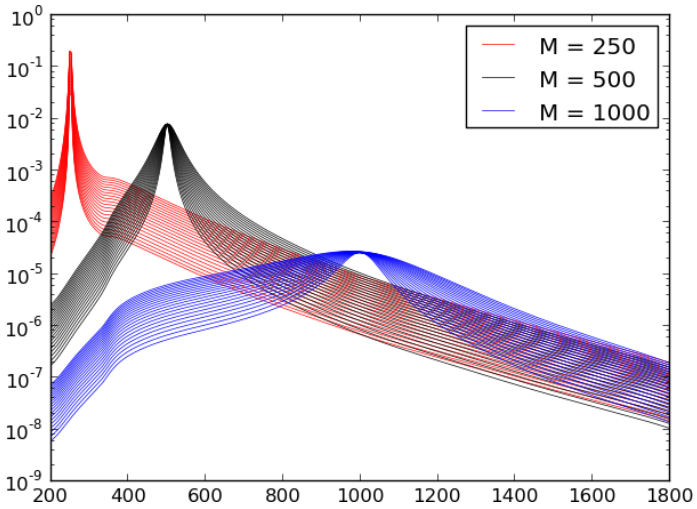


Figure 11.4: The interpolated $|S|^2_{LO}$ shapes for $m_H = 250, 500$ and 1000 GeV and varying values of BSM coupling modifier $C \in [0.5, 1.0]$.

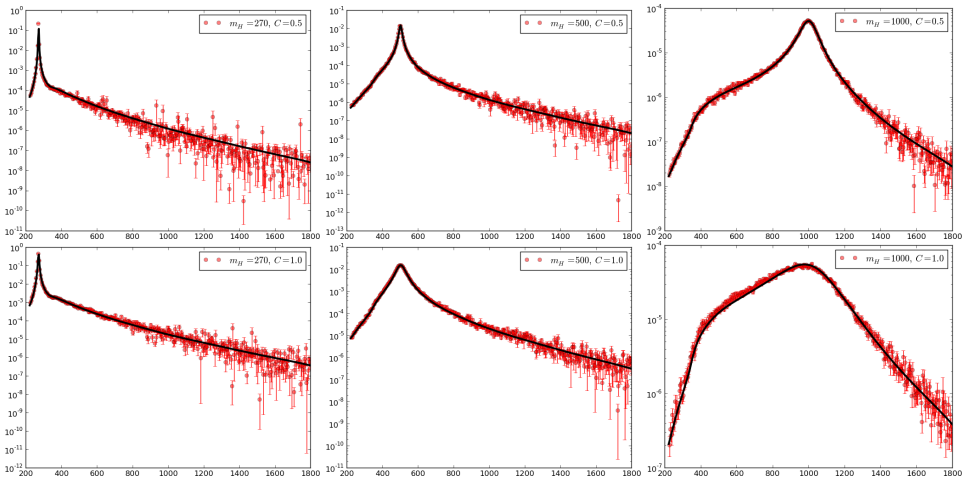


Figure 11.5: Comparison of predicted (black line) and `gg2VV`-simulated (red dots) shapes for $|S|^2_{LO}$.

procedure, one obtains three alternative estimates for signal + interference:

$$\left(|S|^2 + I\right)^{add}_{NNLO} = K \cdot |S|^2_{LO} + I_{LO}, \tag{11.2a}$$

$$\left(|S|^2 + I\right)^{central}_{NNLO} = K \cdot |S|^2_{LO} + \sqrt{K_{gg}} \cdot I_{LO}, \tag{11.2b}$$

$$\left(|S|^2 + I\right)^{mult}_{NNLO} = K \cdot |S|^2_{LO} + K \cdot I_{LO}. \tag{11.2c}$$

The additive $\left(|S|^2 + I\right)^{add}_{NNLO}$ and multiplicative $\left(|S|^2 + I\right)^{mult}_{NNLO}$ estimates are obtained using the so-called K-factor:

$$K = \frac{|S|^2_{NNLO}}{|S|^2_{LO}}. \tag{11.3}$$

The two estimates, however, do not provide a cancellation between signal and background, required by the unitarity in the $m_{ZZ} \to \infty$ limit. The "central" lineshape $\left(|S|^2 + I\right)^{central}_{NNLO}$ does not have this disadvantage. It is defined by introducing another K-factor with only gluon-gluon initial state contributing to the NNLO processes:

$$K_{gg} = \frac{|S|^2_{NNLO}(gg \to H(g) \to ZZ(g))}{|S|^2_{LO}(gg \to H \to ZZ)}. \tag{11.4}$$

It is recommended in [3] to use the central shape (11.2b) as the pivotal lineshape estimate. The additive (11.2a) and multiplicative (11.2c) shapes should be used to set the uncertainty on the approximation.

In order to explore the parameter space of the EWK singlet model one needs to obtain the NNLO signal+interference lineshape as a function of values of m_H and C. Obtaining the distributions on a fine $m_H \otimes C$ grid using `gg2VV` is a very computationally demanding task, that also potentially suffers from low statistics in phase space regions away from the resonance. Having a way to parametrize and interpolate between the lineshapes can be beneficial for the BSM heavy Higgs analyses.

The LO signal is modeled using a product of the relativistic Breit-Wigner function and a common factor $G(m_{ZZ})$ that is independent on m_H and C:

$$f_s(m_{ZZ}|m_H, C) = \frac{m_{ZZ}}{(m^2_{ZZ} - m^2_H)^2 + m^2_{ZZ} \cdot \Gamma^2} \cdot G(m_{zz}), \tag{11.5}$$

where the width parameter of the Breit-Wigner function Γ is allowed to float in the fit. The function (11.5) is fitted to different distributions, produced with

gg2VV for a number of points in $m_H \otimes C$ parameter space. The fitted values of $\Gamma(m_H, C)$ are then interpolated using a RBF spline function, as demonstrated on Figure 11.2. The common factor $G(m_{ZZ})$ in (11.5) is shown on Figure 11.3. It is obtained by first fitting all the signal shapes with simple Breit-Wigner, dividing the shape with the fit and averaging over all the obtained factors.

The resulting interpolated distribution $f_s(m_{ZZ}|m_H, C)$ is demonstrated on Figure 11.4. Comparison of the predicted lineshapes and gg2VV simulated $|S|^2_{LO}$ distributions for a number of $m_H \otimes C$ points is shown on Figure 11.5.

Interference contribution I_{LO} is modeled with the following function, motivated by the so-called "complex mass scheme" [5, 6]:

$$f_i(m_{ZZ}|m_H, C) = Re\left(\frac{e^{-\alpha m_{ZZ}}}{m^2_{ZZ} - m^2_H + i m_H \Gamma}\right). \tag{11.6}$$

The two parameters α and Γ are fitted and interpolated according to a procedure, similar to the fitting procedure for the $|S|^2_{LO}$ contribution. The comparisons of the fit and simulated I_{LO} distributions are shown on Figure 11.1.2

Finally, the resulting interpolated $(S+I)^{central}_{NNLO}$ lineshapes, obtained using (11.2b) are demonstrated on the Figure 11.1.2 with thin lines demonstrating smooth transition between the shapes as the Higgs mass m_H changes. The other two plots on Figure 11.7 show the obtained uncertainty bands for $m_H = 300$ and $m_H = 600$ GeV. The bands are bounded by $(S+I)^{mul}_{NNLO}$ and $(S+I)^{add}_{NNLO}$.

The python software package, that provides all the interpolated shapes discussed above, can be downloaded from [7].

11.2 2HDM

Being one of the most favorable extensions of the simplest standard model (see Section 3.3), the Two Higgs Doublet Model provides a candidate for a heavy neutral scalar particle. So, in principle, it might be possible to constrain the 2HDM parameters using the results of "SM-like scalar" searches.

The vast parameter space of the 2HDM requires some careful planning of the exclusion strategy. As it was discussed in Section 3.3, the model has 14 real parameters in its most general formulation. We first reduce the parameter space by considering the softly Z_2-symmetric 2HDM ($\lambda_6 = \lambda_7 = 0$ in

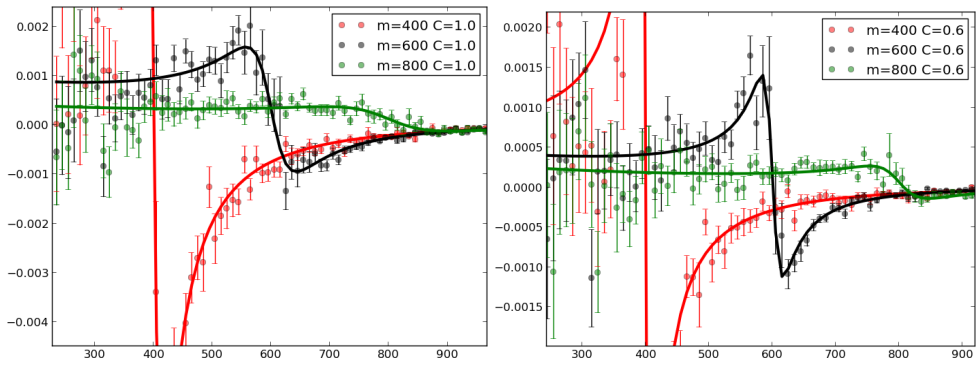


Figure 11.6: I_{LO} modeling/simulation comparison for $m_H = 400, 600$ and 800 GeV and $C = 0.6, 1.0$.

(3.15)) without explicit or spontaneous CP-violation (all parameters are real and $\xi = 0$ in (3.17)). Such a model has 8 free parameters.

By using the *Higgs basis* discussed in the Section 3.3.2 we can further constrain the paramteriztaion by setting a couple of already known physical observables: $m_h = 126$ GeV and $v^2 = 246^2\,\text{GeV}^2$. We are, thus, left with six parameters defining the benchmark model:

$$m_H,\ m_A,\ m_{H^\pm},\ m_{12},\ tan\beta,\ \text{and}\ sin(\beta - \alpha),$$

with α being a mixing angle between the observed SM Higgs h and the heavier scalar H. Further investigation of the six-dimensional parameter space is done by performing a parameter scan.

11.2.1 2HDM parameter scans and results presentation

The 2HDM paramter scan was performed on the following regular grid in the 6D parameter space:

- Yukawa Model I, Model II
- $sin(\beta - \alpha)$ 0.5 – 1.0, step 0.0125
- $tan\beta$ 0.1 – 3.0, step 0.2
- m_{12} 0 GeV – 600 GeV , step 50 GeV
- $m_{H^\pm}$ 300 GeV – 1000 GeV , step 100

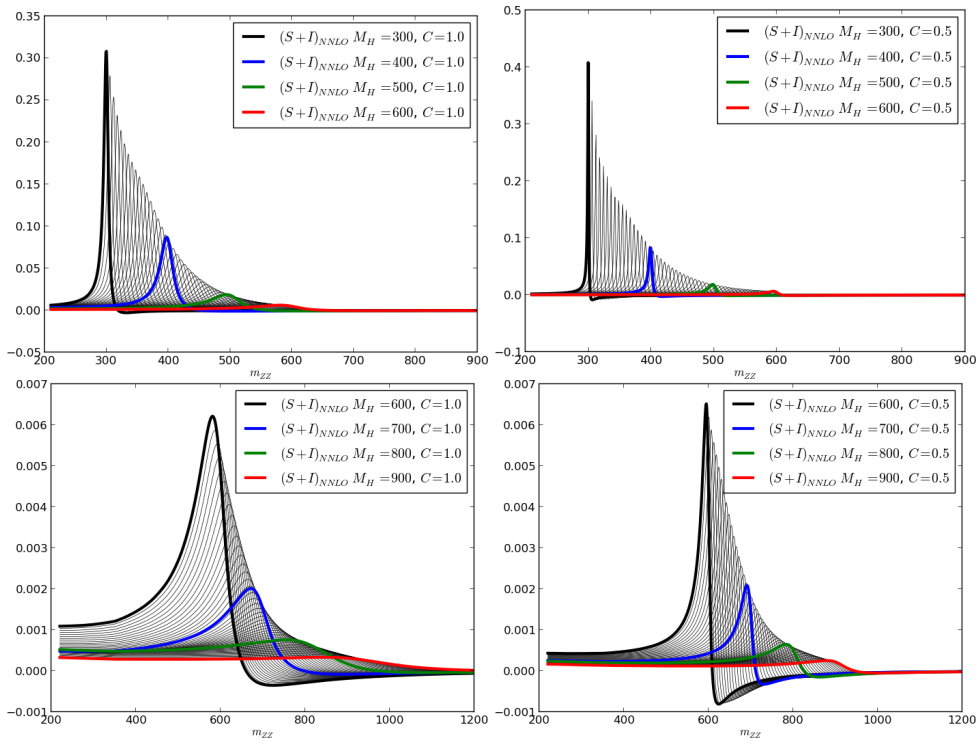


Figure 11.7: Interpolated central lineshapes for $C = 0.5, 1.0$ and running values of m_H.

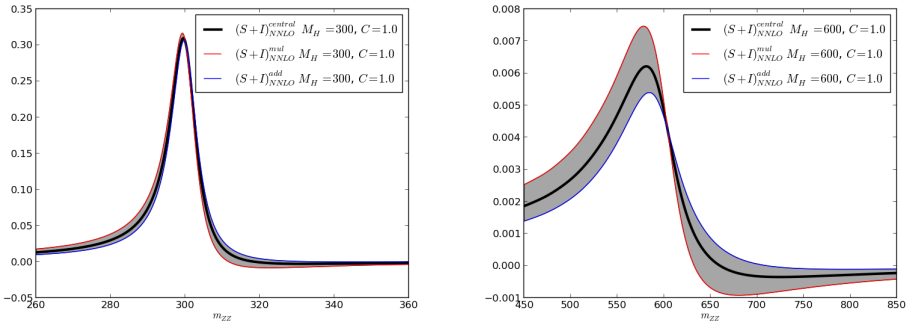


Figure 11.8: The lineshape uncertainty bands around the central shape for $m_H = 300$ GeV (left) and for $m_H = 600$ GeV (right).

- m_A 30 GeV – 1000 GeV , step 100
- m_H 200 GeV – 1000 GeV , step 50

Each point in the parameter space was tested if it satisfies the theoretical constraints – unitarity, perturbativity and vacuum stability by means of the `2HDMC` package [8]. Among the tested $\sim$ 6000000 points only $\sim$ 280000 passed the theoretical constraints. `2HDMC` also provided the h and H scalars branching ratios and widths for the theoretically sound parameter sets.

For calculation of the production cross-sections we've used `SusHi` software [9], interfaced with the `2HDMC` package. For all the points in theoretically acceptable parameter space, the $gg \to h$ and $gg \to H$ cross-sections was calculated. This required several months of CPU time, so the computations was performed by adapting `SusHi` to run on `GRID` [10].

The recorded values of the cross-sections and branchings ratios in the scan can be obtained at [11] and browsed using a web interface at [12].

Bibliography

[1] CMS Collaboration. "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$ = 7 and 8 TeV". *JHEP* 1306 (2013), p. 081. arXiv:1303.4571 [hep-ex].

[2] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Phys.Lett.* B716 (2012), pp. 1–29. arXiv:1207.7214 [hep-ex].

[3] Giampiero Passarino. "Higgs Interference Effects in $gg \to ZZ$ and their Uncertainty". *JHEP* 1208 (2012), p. 146. arXiv:1206.3824 [hep-ph].

[4] Nikolas Kauer and Giampiero Passarino. "Inadequacy of zero-width approximation for a light Higgs boson signal". *JHEP* 1208 (2012), p. 116. arXiv:1206.4803 [hep-ph].

[5] Nikolas Kauer. "Interference effects for $H \to WW/ZZ \to \ell\bar{\nu}u_\ell\bar{\ell}\nu_\ell$ searches in gluon fusion at the LHC" (2013). arXiv:1310.7011 [hep-ph].

[6] Ansgar Denner et al. "Predictions for all processes $e^+e^- \to 4\text{fermions} + \gamma$". *Nucl.Phys.* B560 (1999), pp. 33–65. arXiv:hep-ph/9904472 [hep-ph].

[7] Konstantin Kanishchev. "https://github.com/KKostya/HiggsLineshape" (2013).

[8] David Eriksson, Johan Rathsman, and Oscar Stal. "2HDMC: Two-Higgs-Doublet Model Calculator Physics and Manual". *Comput.Phys.Commun.* 181 (2010), pp. 189–205. arXiv:0902.0851 [hep-ph].

[9] Robert V. Harlander, Stefan Liebler, and Hendrik Mantler. "SusHi: A program for the calculation of Higgs production in gluon fusion and bottom-quark annihilation in the Standard Model and the MSSM". *Computer Physics Communications* 184 (2013), pp. 1605–1617. arXiv:1212.3249 [hep-ph].

[10] E.W. Vaandering. "CMS distributed analysis infrastructure and operations: Experience with the first LHC data". *J.Phys.Conf.Ser.* 331 (2011), p. 072035.

[11] Konstantin Kanishchev. "`https://twiki.cern.ch/twiki/bin/view/Sandbox/CERNonly/THDMParameters`" (2013).

[12] Konstantin Kanishchev. "`http://cern.ch/kkanishc/2HDM/home.shtml`" (2013).

Summary

This doctoral thesis describes the efforts to observe Beyond Standard Model Higgs boson in high energy proton collisions at the Large Hadron Collider with the CMS detector.

The search of the MSSM Higgs boson in the $pp \to bbH \to bbbb$ processes in the semileptonic channel was performed with the data, collected during 2011 at $\sqrt{s} = 7$ TeV proton collisions with total integrated luminosity of $4.8\, fb^{-1}$. No significant excess with respect to the expected SM multijet QCD background was observed. The background was predicted by means of two independent data driven methods – the b-tagging matrix method and the "Hyperball" algorithm. The results, combined with full-hadronic channel was interpreted as the exclusions on the $m_A \otimes \tan\beta$ parameter space, covering a large region of parameters that was not previously rejected by MSSM searches.

The search of the SM-like Higgs boson, decaying into a pair of Z-bosons, which subsequently decay into a pair of leptons and into a pair of quarks, $H \to ZZ \to 2\ell 2q$, was done using the 2012 data corresponding to an integrated luminosity of $19.6\, fb^{-1}$ at proton-proton collisions at center-of-mass energy of $\sqrt{s} = 8$ TeV. The studied Higgs mass hypothesis m_H span the range from 200 GeV to 650 GeV. The results are combined with the previous search at $\sqrt{s} = 7$ TeV and with searches in other final states: 4ℓ, $2\ell 2\ell'$, $2\ell 2\tau$ and $2\ell 2\nu$. Combination of these results allow us to reject the SM-like Higgs with masses up to ~ 1 TeV.

The ongoing work on the reinterpretaions of these searches in a context of two simple BSM extensions is also reported. For the electroweak singlet model, the simple signal lineshape function was developed – it can be used to reweight the signal Monte Carlos getting exclusions for this model parameter space. For the 2HDM, the scan over the six-dimentional parameter space

was performed – it could be used to select most relevant parameters and representative benchmark points in the vast parameter space of the 2HDM.